Vietnamese compounds show an anti-frequency effect in visual lexical decision

Hien Pham[1,2] & Harald Baayen[3,4]

[1]Vietnam Academy of Social Sciences, Vietnam & [2] Vietnam National University, Hanoi &
[3]University of Alberta, Canada & [4]University of Tübingen, Germany

Running Head: Lexical processing in Vietnamese

Corresponding author:
Hien Pham
Institute of Lexicography and Encyclopedia
Vietnam Academy of Social Sciences
36 Hang Chuoi
Hai Ba Trung, Hanoi
Vietnam
e-mail: hpham@ualberta.ca,
phamhieniol@gmail.com

Abstract

Although Vietnamese has a long history of linguistic research, as yet no psycholinguistic studies addressing lexical processing in this language have been carried out. This paper is the first to investigate lexical processing in Vietnamese, and addresses the reading of Vietnamese bi-syllabic compound words. A large single-subject experiment with 20,000 words was complemented by a smaller multiple-subject experiment with 550 words. We report the novel finding of an inhibitory, anti-frequency effect of Vietnamese compounds' constituents. We show that this anti-frequency effect is predicted by a computational model of lexical processing grounded in naive discrimination learning. We also show that predictors derived from this model provide a much better fit to the observed reaction times than traditional lexical distributional predictors. Effects of the density of the compound graph, previously observed for English were replicated for Vietnamese. Furthermore, tone diacritics were found to be important predictors of silent reading, providing further evidence for the role of phonology in reading.

Keywords: compounds, Vietnamese, generalized additive modeling, shortest path lengths, naive discriminative learning

Introduction

Vietnamese is famous as a textbook example of a morphologically isolating language (Lyons, 1968), a language with no morphology. According to (Anderson, 1985, p. 8), Vietnamese is a language "with nearly every word made up of one and only one formative (indeed, one syllable)", (see also Nguyễn, 1996, 2011). The goal of this paper is to show that Anderson's (and Nguyen's) characterization may be both correct and incorrect. It is incorrect for the simple reason that in a lexical database of Vietnamese constructed by the first author, of a total of 28412 words, no less than 22705 (80%) are words that to all practical purposes resemble compounds as familiar from English. For instance, *tàu hoả* 'train', contains the words *tàu*, 'ship', and *hoả* 'fire', and *tàu bay* 'aircraft', contains the word *tàu* 'ship', and *bay* 'fly', just like English *fire engine* contains the words *fire* and *engine*. It is true that Vietnamese has no inflection nor any derivation, but it is rich in compounds. And yet, we shall see that in reading, these compounds are far more like morphologically simple words than English compounds.

Vietnamese (tiếng Việt), spoken by approximately 90 million people, belongs to the Việt-Mường sub-branch of the Vietic branch of the Mon-Khmer family, which is itself a part of the Austro-Asiatic family. In this tone language, all syllables are single morphemes and all morphemes are monosyllabic. Vietnamese linguists have introduced the term *syllabeme* to refer to the syllable-morpheme identity (see e.g., Ngô, 1984, for further information on syllabeme), and we adopt their terminology in this study. Vietnamese words may consist of one syllabeme (e.g., *cây* 'tree', *gạo* 'rice', *mắt* 'eye') or multiple syllabemes, e.g., *hoa hồng* 'rose' (lit. flower pink), and *tàu hoả* 'train' (lit. ship fire).

In the present-day alphabetic writing system of Vietnamese, a syllabeme is written as a sequence of Roman letters, with additional diacritics for distinguishing phonemes that are not properly distinguished by the Roman alphabet, and with additional diacritics for the tones of Vietnamese (*ngang* mid level, *huyền* low falling (breathy), *hỏi* mid falling (-rising), harsh, *ngã* mid rising, glottalized, *sắc* mid rising, tense, and *nặng* mid falling, glottalized, short). Syllabemes are separated by spaces. This spacing convention follows that of its neighbor China, albeit without using the characters familiar from this country's orthography. The result is a straightforward writing system that enables Vietnamese speakers to learn how to read and write within a few months. It serves as the official orthography nation-wide (Nguyễn, 1997).

Vietnamese syllables are phonotactically severely restricted, and consist of an optional onset consonant, followed optionally by a bilabial consonant glide, followed by an obligatory vowel (with one of six tones), followed optionally by a single coda consonant. Table 1 presents a partition of the most common syllabemes in contemporary Vietnamese. The total number of attested syllabemes in actual use is 6,651, with a syllabeme type defined as a unique character sequence between spaces. By comparison, the total number of English syllables as attested in the CELEX lexical database for English wordforms (Baayen et al., 1995), differentiated for stress (no stress, primary stress, secondary stress) is 17,918. Without differentiating between stress, the number of different syllables remains substantially larger than in Vietnamese (11,492).

Although almost all syllabemes are independent words, the majority of words in Vietnamese comprise more than one syllabeme. Two-syllabeme compounds often show the same lack of semantic transparency that characterizes compounds in English. Knowing the meanings of the constituents *ship* and *fire* is not sufficient to deduce the compound's meaning (in Vietnamese: a means of transportation making use of rails, in English: a truck designed for putting out fires).

The combination of a limited set of syllables (compared to English), the conflation of syllables and morphemes, and rampant compounding raises the question of how compounds are processed. Are they read as two-syllable words, or are they processed through some form of morphological decomposition?

Table 1: Vietnamese syllable type frequency

Type	Frequency	Example	English gloss
CwV	141	*hoa, quê*	flower, countryside
CwVC	436	*hoang, xoay*	uncultivated, revolve
wV	11	*oà, uỷ*	burst out crying, commissioner
CV	1106	*ngủ, xu*	sleep, coin
wVC	27	*oách, oằn*	dapper, to curve
CVC	4681	*bên, xương*	side, bone
V	50	*ả, ý*	lass, idea
VC	188	*ác, ai*	fierce, anybody

In what follows, we first introduce a computational model for lexical processing based on naive discriminative learning that predicts for Vietnamese that high-frequency constituents delay comprehension. The same model architecture, applied to English, predicts, in line with many empirical studies on this language, facilitation from constituents with high frequencies and large morphological families. This surprising prediction of the computational model is then tested against two lexical decision experiments, one with a single subject (the first author) reading 20,000 words, and one with multiple subjects reading a smaller subset of 550 words. The first experiment is an exhaustive experimental survey of all two-syllabeme compounds of Vietnamese listed in a major dictionary (Hoàng, 2000). The second experiment is a multiple-subject replication study. We then consider the computational model in further detail, and conclude with a discussion and evaluation section.

Predicting lexical processing in Vietnamese with naive discriminative learning

Naive discriminative learning is a theory of lexical processing which builds on the Rescorla-Wagner equations and the equilibrium equations thereof (Wagner and Rescorla, 1972; Danks, 2003).

Central to this learning theory is how well *cues* discriminate between *outcomes*. By way of a non-linguistic example, consider cues such as `having whiskers`, `having fur`, and `having paws`, for outcomes suchs as RABBITS, MICE, CATS, and PORCUPINE. Consider a picture with a rabbit, with the rabbit's whiskers clearly visible. In this situation, the weight on the link from `having whiskers` to RABBIT is increased, whereas the weight on the link from `having whiskers` to PORCUPINE are decreased. Importantly, the weights from `having whiskers` to MICE and CATS are decreased as well, reflecting that `having whiskers` incorrectly predicted that the picture would be about a a mouse or a cat. This may seem counterintuitive, but it reflects that learning is error-driven (Rescorla, 1988; Marsolek, 2008; Ramscar et al., 2010), a finding for which excellent neurophysiological evidence has been obtained (Schultz, 1998).

Naive discriminative learning (henceforth NDL) applies these insights to language, offering the possibility to estimate how well orthographic *cues* (letters, letter pairs, or letter trigrams) activate lexemic *outcomes*. Here, we use the term *lexeme* in the sense of Aronoff (1994) to denote a representation mediating between form and world knowledge. For the present purposes, the lexemes can be thought of as the symbolic gateways to semantic, pragmatic, and encyclopedic lexical knowledge. NDL is an a-morphous theory: there are no representations for stems, morphemes, or exponents. It is most closely related to Word and Paradigm Morphology (Matthews, 1974; Blevins, 2003) in theoretical linguistics. In short, the model provides estimates of how well simple orthographic cues predict lexemic outcomes.

The model's predictions are derived from corpora or lexical databases. Central to the algorithm is the definition of a learning event. A learning event consists of a set of orthographic cues, such as the

orthographic digraphs {#q, qa, ai, id, d#} (with the hash denoting the space character), and a set with one (or more) lexemes, such as {QAID} (a legal scrabble word meaning tribal chieftain). Given the sets of cues and outcomes, the Rescorla-Wagner equations are applied to update the weights from these orthographic cues present to all lexemes that the model has encountered. Thus, the weight on the link between #q to QAID is strengthened, whereas the weight on the link to QUESTION is weakened.

When applied rigorously to large corpora or databases, NDL correctly predicts a wide range of phenomena in the lexical processing literature (Baayen et al., 2011; Baayen, 2010a, 2011; Baayen et al., 2013; Mulder et al., 2014; Ramscar et al., 2010). For English bi-morphemic compounds, higher frequency constituents afford shorter response latencies. This is mirrored exactly in NDL's predictions for this language (Baayen et al., 2011).

Returning to Vietnamese, in order to evaluate the potential consequences for lexical processing of a lexicon combining productive compounding with a small set of a phonotactically highly constrained syllabemes, we trained an NDL model (using the R code available in the ndl R package, Shaoul et al., 2013) on 27181 words, of which 5471 consisted of one syllabeme, and 21710 contained two syllabemes. Word frequencies ranged from 1 to 1.1552×10^6. We used letter bigrams as cues, and compounds' lexemes as outcomes. For instance, for the compound *tàu hoả*, the model was supplied with the set of letter digraphs (#t, tà, àu, u#, #h, ho, oả, ả#) and the outcome TRAIN. As *tàu hoả* occurred 216 times in our corpus, the model was trained on 216 learning events in which the above letter bigrams were paired with the lexeme TRAIN.

Following (Milin et al., 2014), we estimated the model's support for a given lexeme with the product of the word's activation (the summed weights on the connections of the word's cues in the visual input, to its lexeme) and the median absolute deviation of the weights on all connections feeding into that lexeme (irrespective of whether they are present in the visual input). For the statistical analysis, this product was log-transformed to remove the rightward skew in its distribution. The log-transformed support measure was subjected to a change in sign to obtain a simulated response latency (words with greater support should be responded to with shorter response latencies).

In order to understand how the simulated response latencies relate to standard lexical distributional measures, we compiled a set of 18 (highly correlated) corpus-based counts, serving to predict both the latencies in the experiments reported below, and the latencies simulated by the NDL model. These counts included several measures of frequency of occurrence of the two-syllable words in a newspaper corpus and in a subtitle corpus, as well as measures of dispersion (contextual diversity) in these corpora. Furthermore, corresponding counts were collected for the first and second syllabemes. In addition, the primary (Moscoso del Prado Martín et al., 2004) and secondary (Baayen, 2010b; Mulder et al., 2014) family size counts for the syllabemes were obtained, as well as their dispersion. Finally, additional family size counts were compiled for the constituents, once disregarding only diacritics for tone, and once disregarding all diacritics. For further information on the lexical resources on which these counts are based, see Pham (2014).

As the collinearity of this set of predictors was very high (as indexed by the κ index of collinearity of Belsley et al. (1980), which for our data was 610.58; values above 30 are considered as indicating very severe collinearity), we orthogonalized them using principal components analysis (for an introduction to this method, see, e.g., Baayen, 2008). A screeplot revealed three primary principal components. The first principal component, henceforth `Compound Frequency PC`, revealed large negative loadings for the compound frequency and dispersion measures. Constituent family size measures, with or without diacritics, had reduced negative values on this component. The second principal component contrasted morphological family size measures (large negative loadings) and constituent frequency measures (with somewhat smaller negative loadings) with compound frequency and dispersion measures (large positive loadings). This component is henceforth referred to

as `Part-Whole Balance PC`, as it contrasts words with prominent constituents and low compound frequency with words with high compound frequency and constituents with small family size and frequency. The third principal component, `Positional Family Size PC`, contrasted family size measures for the second syllabic constituent (large negative loadings) with family size measures for the first syllabic constituent (large positive loadings). The proportion of the variance captured by the three principal components were 0.37, 0.23, and 0.18.

A linear regression model fitted to the simulated latencies with the first two principal components as predictors supported a positive slope for `Compound Frequency PC` ($\hat{\beta} = 0.48, p < 0.0001$) and a negative slope for `Part-Whole Balance PC` ($\hat{\beta} = -0.71, p < 0.0001$). Since measures for the frequency of the compound have large negative loadings on `Compound Frequency PC`, the model predicts that more frequent compounds will be responded to more quickly, as expected. Furthermore, since constituent family size and frequency measures have large negative loadings on `Part-Whole Balance PC`, the model predicts that reading is slowed down when the constituent frequencies and family sizes are large. This prediction of interference from constituents with large family sizes and greater frequency for Vietnamese is surprising in the light of the facilitation typically found for lexical decision in English (Baayen et al., 2010, 2011). We therefore now consider two lexical experiments in Vietnamese, in order to ascertain whether the model's prediction of an anti-frequency effect for constituent syllabemes is correct.[1] We first report a large single-subject experiment that covers the full range of items on which the NDL model was trained. We then present a second study with a many participants responding to a small subset of the words in Experiment 1.

Experiment 1: A single-subject large-scale lexical decision experiment

Method

Materials All disyllabic words from the Vietnamese Dictionary (Hoàng, 2000) were selected, with the exception of those words involving reduplication, resulting in a list of target words comprising 15021 words. In addition, nearly 5000 single syllabeme (monomorphemic) words were included, resulting in a total of 20,000 Vietnamese words. (For the importance of comprehensive numbers of items, see, e.g., Balota et al., 2004; Ferrand et al., 2010; Keuleers et al., 2012).

For the statistical modeling of the response latencies, we considered several additional predictors in addition to the three principal components introduced above: the length of the compound (in letters), session number (1–16), the time of day the block was run (in minutes from midnight; the translation into clock time is given at the top of the panel), the lexical tone of the first syllable (1–6) as well as that of the second syllable (1–6), and the word category of the compound. Table 2 presents the distribution of tones.

Table 2: Distribution of tones in Vietnamese single-syllabeme and two-syllabeme words.

Tone	Single Syllabeme		First Compound Syllabeme		Second Compound Syllabeme	
	types	tokens	types	tokens	types	tokens
ngang	984	14,130,780	6641	5,059,200	4693	3,443,209
huyền	802	11,543,156	3840	2,586,797	3360	2,295,111
ngã	313	3,314,686	858	386,988	1054	547,700
hỏi	514	5,075,897	2145	1,884,127	2277	1,868,108
sắc	1365	11,823,632	5507	4,128,831	5918	4,015,755
nặng	976	7,218,239	3361	2,784,402	4995	4,560,463

As fixed-effect factors we included whether the first/second syllable constituents are also used as classifiers, and whether the compound is part of a strongly connected component of the Vietnamese directed compound graph. A strongly connected component of a directed graph is a subgraph with the property that each vertex (node) in the graph can be reached from any other vertex by following the directed edges (links). Baayen (2010b) studied the directed compound graph of English (restricted to bi-morphemic compounds), i.e., a graph in which compound constituents are the vertices, and in which directed edges connect first constituents to second constituents. The English compound graph has one (large) strongly connected component. The Vietnamese compound graph is characterized by two (also large) strongly connected components. Compounds in a strongly connected component are part of a particularly dense area of the lexicon. Just as neighborhood density at the segment level (Chen and Mirman, 2012; Balota et al., 2004) may affect lexical processing, neighborhood density at the syllabeme/constituent level may help explain response latencies.

Within a strongly connected component, cyclic chains exist, as illustrated in Figure 1. In this graph, each pair of nodes linked by a directed edge represents an existing compound, with constituents ordered as indicated by the direction of the arrows. A numeric predictor that comes into play only for words in the strongly connected component is the length of the shortest path from second syllabeme to the first. In Figure 1, these shortest path lengths are 2, 4, 8, and 10 respectively.

For each of the 20,000 words in the experiment, a pseudoword was generated using the Wuggy pseudoword generator (Keuleers and Brysbaert, 2010). Each pseudoword differed from its reference word by one subsyllabic segment (i.e., the onset, nucleus, or coda) per syllable. As a consequence, a two-syllable nonword differed in two positions from its reference word. A further constraint on pseudoword generation was that the position selected for change was chosen such that it resulted in the smallest possible overall change in syllable frequency, transitional frequency between syllables, and subsyllabic frequency. As a result, the pseudo-morphological structure of the nonwords resembled the morphological structure of the words as closely as possible, as can be seen in Table 3. The distribution of tone diacritics in the nonwords also faithfully reflected their distribution in existing words.

Table 3: Examples of compound words and their equivalent pseudowords. None of the pseudowords are existing word in Vietnamese.

Word	Pseudoword
ác cảm	ác bạm
á hậu	á đấu
ẩn nấp	ẩm bấp
âm hưởng	âm bượng
áp thấp	áp cháp
nghị sĩ	nghì sự
thể nghiệm	thử nghiêm
vị thế	vù thị
xoắn ốc	xoán óc
xuất viện	xuất tiên

Subject The first author, a native speaker of Vietnamese, served as the single participant of this experiment. Responding to all forty thousand trials required 46 hours, over a 4-week period.
Procedure All the stimuli, including both words and nonwords, were merged into one list. A script was written to randomly select equal numbers of word and pseudoword stimuli from the list, which were

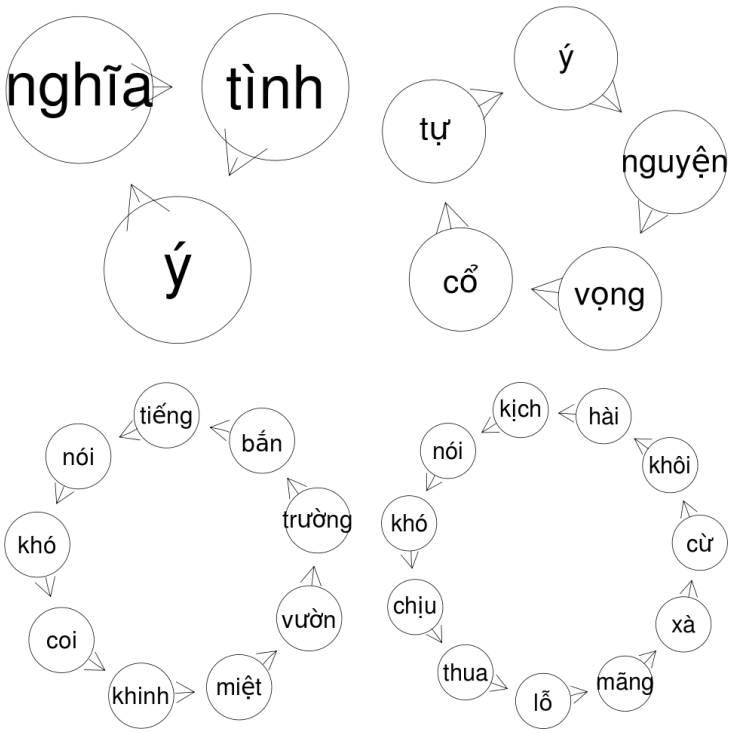

Figure 1: Examples of cycles in the compound directed graph: shortest head-to-modifier paths for *ý* → *nghĩa*, *ý* → *nguyện*, *miệt* → *vườn*, and *xà* → *cừ*. English glosses of the compounds for the upper left panel: *nghĩa tình* 'sentimetal attachment', *tình ý* 'intention', *ý nghĩa* 'mean, sense'; for the upper right panel: *ý nguyện* 'wishes', *nguyện vọng* 'aspiration', *vọng cổ* 'name of a traditional tune', *cổ tự* 'ancient writing', *tự ý* 'willingly'; for the lower right panel: *kịch nói* 'play', *nói khó* 'beg', *khó chịu* 'uncomfortable', *chịu thua* 'yield', *thua lỗ* 'lose', *lỗ mãng* 'coarse', *mãng xà* 'python', *xà cừ* 'conch, nacre', *cừ khôi* 'splendid', *khôi hài* 'funny, humorous', *hài kịch* 'comedy'; for the lower left panel: *tiếng nói* 'voice', *nói khó* 'beg', *khó coi* 'unsightly, unaesthetic', *coi khinh* 'despise', *khinh miệt* 'despise, think little and scorn', *miệt vườn* 'hick', *vườn trường* 'school garden', *trường bắn* 'rifle range', *bắn tiếng* 'spread word'.

then merged into a template script for DMDX. Thanks to this automated procedure, the participant (who also implemented the experiment) remained completely uninformed about the words to appear in a given experimental session. The total experiment comprised 80 blocks of 500 stimuli. Each block took about 60 minutes to finish (including breaks) and was subdivided into five sub-blocks of 100 stimuli each. Between each sub-block, the participant was asked to press the space bar to continue. The participant felt that the interruptions increased his control and provided him with information about his progress through the block. The participant completed a maximum of two blocks per day.

Stimuli were presented on a 17-in. Acer laptop with a refresh rate of 85 Hz and a resolution of 1,600 x 900 pixels, which was controlled by an Intel Core i7 1.6GHz processor. Stimuli were presented in lowercase 26-point Courier New font, and appeared as black characters on a grey background. Stimuli were presented and responses collected with the DMDX software (Forster and Forster, 2003).

The participant indicated as quickly and as accurately as possible whether a presented letter string formed a word or not in Vietnamese by pressing a button on a Microsoft USB wired Xbox 360 game controller for Windows with his left (No) and right (Yes) index fingers. Each trial started

with a centered fixation point '+' that was presented for 500 msec, followed by the target letter string, which stayed on the screen until the participant responded or until 2 seconds had elapsed. The lexical decision experiment started with 12 practice trials in each session, followed by 500 experimental trials, separated by four breaks.

Results

Response latencies were subjected to a scaled negative reciprocal transform (-1000/RT) to reduce the skew in their distribution. In order to properly model nonlinear functional relations in two or more dimensions, we made use of generalized additive mixed-effects regression models GAMMs, (see, e.g., Hastie and Tibshirani, 1990; Wood, 2006) as implemented in the `mgcv` package (Wood, 2006, 2011) (version 1.8.3) of the `R` statistical computing software (R Core Team, 2014).

Generalized additive mixed models extend the standard linear mixed model with tools for modeling nonlinear functional relations between one or more predictors and the response variable. When the relation between the response and a single predictor is non-linear (as, for instance, is the case for the dilation of the pupil as a function of time: the pupil first widens, and then narrows), a thin plate regression spline is the optimal choice. A thin plate regression spline is nothing more than a weighted sum of mathematically simple functions, the so-called basis functions, with a penalty for wiggliness to avoid overfitting. When a response depends on two predictors in a non-linear way, a tensor product smooth can be used to fit a wiggly surface to the data. Just as thin plate regression splines, tensor product smooths are penalized to avoid overfitting. Tensor product smooths provide an important extension of the multiplicative interaction of two (or more) numeric predictors in the linear mixed model. For two predictors, a multiplicative interaction fits a hyperbolic plane to the data, such that when the value of one predictor is fixed, the effect of the other predictor is strictly linear. Although some interactions may be well-described by a multiplicative interaction, many are not — consider, for instance, an "egg-box" like regression surface. The linearity assumption of the standard mixed model often fails to do justice to the actual patterns in the data, and may result in important effects remaining unobserved. Given that previous studies on lexical processing have observed interactions between frequential predictors (typically modeled with multiplicative interactions, see, e.g., Colé et al., 1997; Kuperman et al., 2008, 2009; Miwa et al., 2014) and given improved model fits obtained for such interactions when exchanging linear mixed models for GAMMs (Baayen et al., 2010), we make use of GAMMs in order to obtain an optimal understanding of the quantitative structure of our data.[2]

Tables 4 and 5 summarize the generalized additive mixed model fitted to the inverse-transformed response latencies. First consider the parametric part of the model, summarized in the upper half of Table 4. We find here the regression coefficients, their standard error, and associated t and p values, familiar from standard linear regression models. The positive coefficient for `Word Length` ($\hat{\beta} = 0.016$) indicates that, as expected, longer words tended to elicit longer latencies. The non-significant negative coefficient for words in the strongly connected component of the compound graph (SCC=TRUE, $\hat{\beta} = -0.065$) is suggestive, albeit no more than that, of words that are well-embedded in the lexicon being responded to more quickly.

The second half of Table 4 lists the smooths and random effects in the model. Here, *edf* signifies the effective degrees of freedom, which is roughly the number of parameters invested in a smooth (or random effect). An edf close to 1 for a smooth is indicative of a straight line (which requires one parameter, the slope, in addition to the intercept). The smooth terms of the model are best understood through visualization, presented in Figure 2.

A nearly linear effect of `Frequency PC` indicates that more frequent words, which have more negative scores on this principal component, are responded to faster, as expected (upper left panel).

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.5829	0.0477	-33.1898	< 0.0001
Word Length	0.0160	0.0014	11.0923	< 0.0001
SCC=True	-0.0651	0.0352	-1.8486	0.0645
B. smooth terms	edf	Ref.df	F-value	p-value
smooth Frequency PC	4.0473	5.0885	277.9798	< 0.0001
smooth Part-Whole Balance PC : SCC=False	1.0000	1.0000	207.1894	< 0.0001
smooth Part-Whole Balance PC : SCC=True	3.8749	4.8666	160.2241	< 0.0001
smooth Positional Family Size PC	3.5894	4.5488	3.6806	0.0038
random effect Tone of First Syllable	4.0966	5.0000	7.2894	< 0.0001
random effect Tone of Second Syllable	4.1705	5.0000	4.9090	0.0001
random effect Word Category	7.7133	10.0000	11.7848	< 0.0001
smooth Minutes	4.3712	5.0122	38.3893	< 0.0001
smooth Session Number	8.4037	8.7715	41.3732	< 0.0001

Table 4: Generalized Additive Model fitted to the negative reciprocal transformed lexical decision latencies of the large single-subject study (edf: estimated degrees of freedom); SCC: the factor specifying whether the compound is part of the strongly connected component of the compound graph

The next two panels present the effect of the `Part-Whole Balance PC`, which entered into an interaction with membership in the strongly connected component. The effect of `Part-Whole Balance PC` was linear for words outside the SCC, whereas it was slightly nonlinear for words that are part of the SCC. Comparing the third panel with the second, we find that the effect of the `Part-Whole Balance PC` was stronger for words belonging to the SCC. When the syllabemes of a compound have larger families, and when these families belong to highly interconnected sections of the compound graph, response latencies apparently become progressively longer. (For completeness, we note that when separate predictors for constituent frequencies are considered, they likewise give rise to inhibitory effects; models not shown.)

The fourth panel indicates a modest somewhat U-shaped effect for `Positional Family Size PC`. Recall that large negative values on this principal component reflect large families for the second syllable, whereas large positive values reflect large families for the first syllable. Apparently, when the families are out of balance, i.e, when the one family is large at the expense of the other, then responses are delayed. Processing appears to be optimal when both families are in balance (i.e., when `Positional Family Size PC` assumes values around zero). A similar trade-off was observed by DeCat et al. (2014a,b) in the EEG elicited by English compounds.

Table 4 indicates that all three random-effect factors (the tone on the first syllable, the tone on the second syllable, and word category) contribute significantly to the model fit (all $p < 0.0001$). The coefficients for these random effects factors are shown in panels 5 through 7 by means of quantile-quantile plots. We incorporated these predictors as random-effect factors instead of as fixed-effect factors for several reasons. First, this helps us avoid tables of coefficients that are cluttered with many contrast-coefficients that only represent a subset of the possible contrasts between the many group means of these multi-levelled factors. Second, for these factors, we do not have any a-priori hypotheses as to what levels should differ. We include these predictors because we predicted them to capture a significant part of the variance, which indeed they do. Fixed-effect coefficients are not of interest to us at this exploratory stage of investigation, because they are less informative. Third, since the coefficients obtained for random-effect factors are shrinkage estimates, we are protected agains overfitting the model.[3]

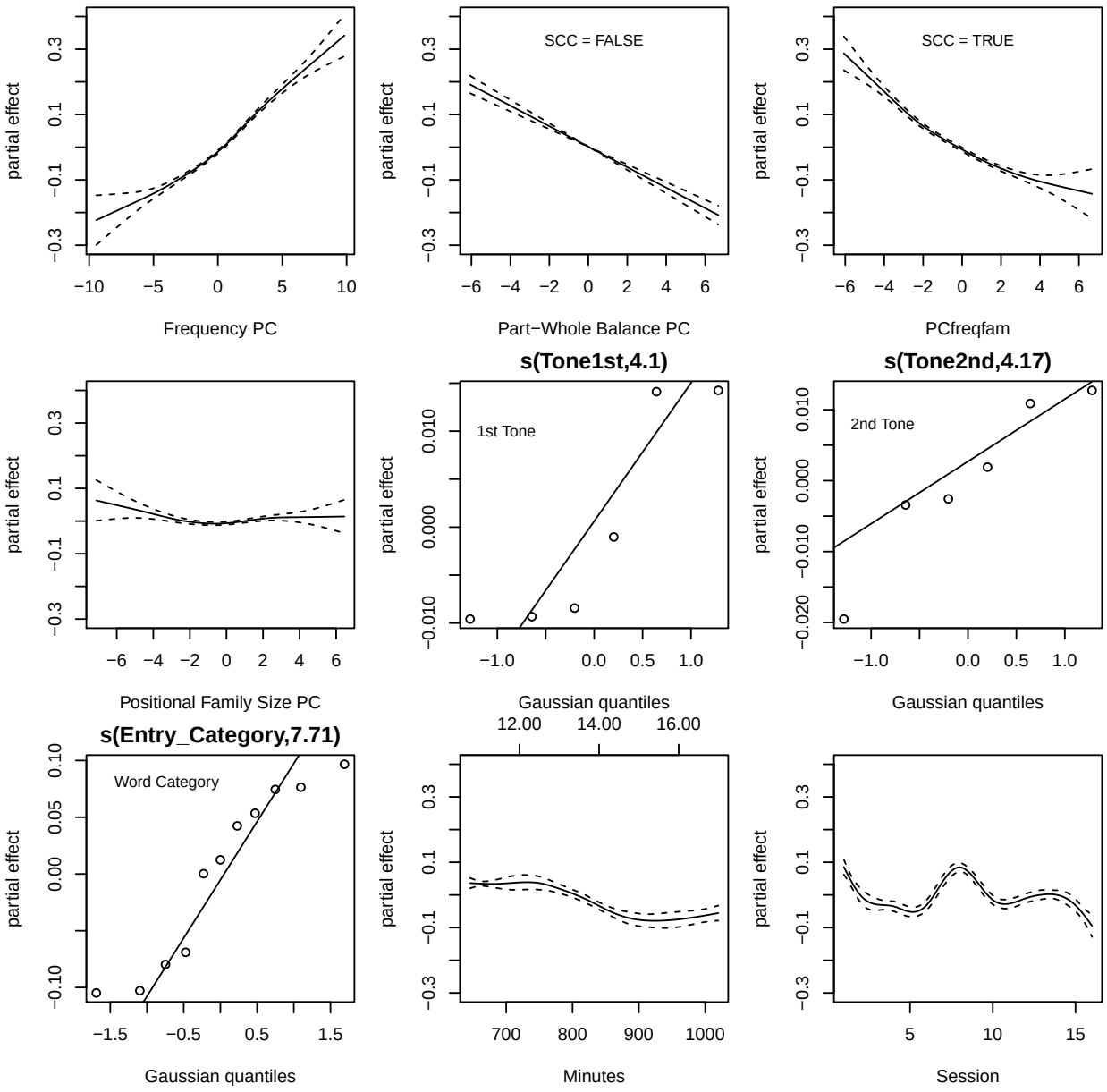

Figure 2: The partial effects of smooths and random effect factors in the model fitted to the negative reciprocal transformed response latencies in Experiment 1. SCC denotes the factor specifying membership in a strongly connected component of the Vietnamese compound graph.

Inspection of the coefficients for the tone of the first syllabeme shows that the *huyền* low falling (breathy) and *sắc* mid rising, tense tones elicited longer latencies than the other four tones. With respect to the second syllabeme, the *ngã* mid rising, glottalized tone elicited the shortest latencies, and the *huyền* low falling (breathy) and *ngang* mid level tone the longest. The major word categories (noun, verb, adjective) were responded to more quickly than the minor word categories.

The last two panels of Figure 2 presents smooths for the time of day at which the experiment was run (`Minutes`) and session number (`Session`). The plot for `Minutes` shows that responses were faster in the afternoon than in the morning. The plot for `Session` indicates that in the course of

this month-long experiment, responses were elongated at the beginning and halfway through the experiment, and that towards the end of the experiment, responses were shorter. We were not able to find any interactions involving these two predictors that would improve the model fit. We also could not detect any further effect of `Trial` (the rank of an item in its experimental list).

models	AIC
+ Minutes and Session	583.09
+ Tone1 and Tone2	90.40
+ Word Category	19.39
+ Word Length	44.12
+ Frequency PC	1817.62
+ Part-Whole Balance PC * SCC	946.48
+ Positional Family Size PC	10.94

Table 5: Reduction in AIC as predictors are added to an intercept only baseline model for the single-subject dataset. SCC: factor indicating membership in the strongly connected component of the compound graph.

Table 5 lists the decrease in AIC[4] when, starting with an intercept-only model, predictors or groups of predictors, are added to the model formula. The most important predictor is `Frequency PC`, unsurprisingly, as it captures the word frequency effect. The second most important predictor is `Part-Whole Balance PC`, which contrasts words with large families and low frequencies with high-frequency words with small families. Next in importance are the experimental variables MINUTES and SESSION. As expected for a language rich in tones, the two tone random effect factors also contribute substantially to the goodness of fit. Contributions of the remaining predictors were modest.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.6509	0.0393	-41.9976	< 0.0001
Word Length	0.0161	0.0017	9.6701	< 0.0001
Second Syl. is Classifier: TRUE	-0.0115	0.0179	-0.6403	0.5220
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
smooth Frequency PC	3.3490	4.2718	167.2776	< 0.0001
smooth Part-Whole Balance PC	3.8493	4.8373	152.8841	< 0.0001
smooth Minutes	3.8974	4.5590	29.4380	< 0.0001
random intercepts tone of first syllable	3.9878	5.0000	4.8267	0.0020
random intercepts tone of second syllable	4.3135	5.0000	5.5958	< 0.0001
random intercepts word category	7.4343	10.0000	7.3111	< 0.0001
smooth Session	8.1698	8.6756	31.4299	< 0.0001
smooth shortest path length	1.0000	1.0000	39.6244	< 0.0001
tensor smooth Sh. Path by Frequency PC : 2nd is Cl = FALSE	2.8869	3.5853	3.0730	0.0199
tensor smooth Sh. Path by Frequency PC : 2nd is Cl = TRUE	1.0000	1.0000	1.1487	0.2838

Table 6: Generalized Additive Model fitted to the negative inverse transformed lexical decision latencies of the large single-subject study, restricted to the words in the strongly connected component of the compound graph (edf: effective degrees of freedom; Cl: classifier).

Table 6 presents a generalized additive mixed model fitted to the subset of compounds that are part of the strongly connected component of the compound graph (11392 of the 15021 observa-

tions). For these compounds, the length of the shortest path from head to modifier is of potential relevance. When the shortest path length is included as predictor, `Positional Family Size PC` loses significance, and interactions emerge with whether the second syllable-constituent is also in use as a classifier. For those compounds with a second constituent that is not also a classifier, and only for these compounds, an interaction of `Frequency PC` by shortest path length was present, as revealed by the tensor product smooth shown in Figure 3. Figure 3 presents the fitted surface as a function of `Shortest Path Length` and `Frequency PC`. Darker colors denote shorter latencies, darker shades of yellow denote longer latencies. As on a terrain map, contour lines connect points that have the same vertical height. Contour lines are 0.05 units apart on the -1000/RT scale.

For this GAMM model, we adopted a decompositional approach with separate smooths for `Shortest Path Length` and `PC freq`, combined with a tensor smooth for the partial effect of the interaction of these two predictors. (Inclusion of the interaction smooths for compounds with second constituents differentiated by their classifier status reduced the AIC by 4.3.) Figure 3 shows that for high-frequency words (large negative values of `PC freq`), the effect of path length is small, with an optimum of shortest responses around paths of length 2–4. As frequency decreases (larger, positive values of `PC freq`), the effect of path length reverses, such that for the lowest frequency words, lengths 4–6 are least optimal, with the longest response latencies. In other words, the word frequency effect is strongest for compounds with a shortest path length of 4–5 — for these two path lengths, the greatest number of contour lines is crossed in Figure 3 when moving horizontally along the Y-axis.

The modulation of shortest path length by frequency is very similar to the interaction of shortest path length by first constituent family size reported in Baayen (2010b) for word naming in English. Interactive activation theories might explain the observed pattern as resulting from activation spreading from the second constituent through the compound graph and ultimately returning to the first constituent, resulting in confusion about the functional status of the first constituent (e.g., modifier in the target compound, but head of the previous compound in the compound chain). This confusion would then arise primarily for low-frequency compounds and intermediate path lengths. For short paths, activation would arrive back too early to interfere, at a time when there still is strong bottom-up support. For long paths, activation would have decayed too much to cause strong interference (see Baayen, 2010b, for further discussion).

Whereas the graph-theoretical effects observed for Vietnamese converge with similar effects observed for English, the sign of the effect of `Part-Whole Balance PC` is different from the empirical record for English. Interestingly, the results for `Frequency PC` and `Part-Whole Balance PC` fit well with the predictions of the NDL model. Apparently, the distributional characteristics of Vietnamese differ such that the same learning model, trained on English, predicts facilitation, whereas when trained on Vietnamese, it predicts inhibition from compounds' constituents. We suspect that the strong phonotactic restrictions on syllabemes are at issue here, resulting in a relatively small set of individually meaningful constituents that are 'recycled' in compounds of varying degrees of transparency, and that are written with intervening spaces. From a discrimination learning perspective, discriminating between the meanings of the constituent syllabemes and the meanings of the compounds is harder in Vietnamese compared to English, because there is more functional overloading of the constituents.

There are some hints in the literature on French, English, and Dutch, that constituents and complex words may be in each other's way. Colé et al. (1997) report, for one of the conditions in one of their experiments, an inhibitory effect of cumulative root frequency for French. Kuperman et al. (2009) observed (using a multiplicative interaction in a linear mixed model) an interaction of left constituent frequency by compound frequency for Dutch. Analyses of response latencies to compounds in the English Lexicon Project (Balota et al., 2004) with generalized additive models also

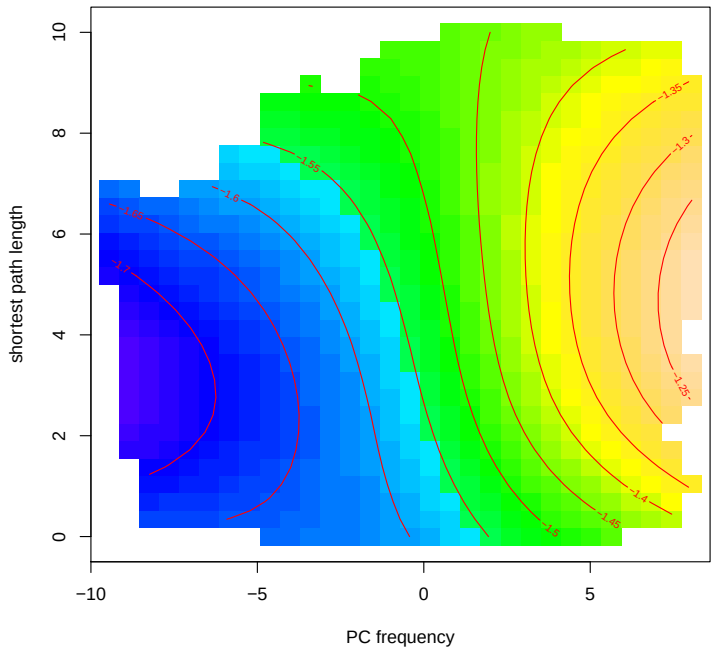

Figure 3: Tensor product surface for the interaction of Shortest Path Length and PC freq for compounds the second constituent of which is not in use as a classifier, in the single subject experiment (Exp. 1).

suggested a (nonlinear) interaction of left constituent frequency by compound frequency, such that for low compound frequencies, very low or very high modifier frequencies resulted in longer lexical decision latencies. None of these studies support the consistent inhibitory effect of high constituent frequency and family size observed for both constituents in Vietnamese compounds.

The empirical results obtained thus far are based on a single subject, albeit on a very large number of words. To further validate the Vietnamese constituent anti-frequency effect, we consider a multiple-subject replication study with a smaller random sample of items.

Experiment 2: Multiple-subject small lexical decision experiment

Experiment 2 was run in Vietnam with 33 participants, and 550 words (and 550 nonwords). The number of items was chosen to provide as extensive coverage as possible within a single experimental session of approximately one hour.

Method

Materials 550 disyllabic compounds were randomly selected from the 15,000 compound items in the single-subject experiment, such that high- and low-frequency compounds had an equal chance of being selected.

Subjects Thirty three students at the Vietnam National University were recruited to take part in the lexical decision experiment (mean age 21.9, range 20 – 22 years, 12 males, 21 females). All participants were native Vietnamese speakers and had at least 14 years of education.

Procedure The same experimental equipment was used as in Experiment 1. Eight lists, each with the items in a different random order, were constructed for counterbalancing; subjects were randomly assigned to these lists. The experiment was administered in the same way as a block in Experiment 1. However, subjects were offerred the possibility of self-timed break after every 100 items.

Results

Table 7 summarizes the generalized additive mixed model fitted to the inverse-transformed response latencies. In addition to the random effect factors for tone and word category, we included random intercepts for item (word). For subjects, we requested a specific kind of random effect, namely, shrunk factor smooths. These factor smooths make it possible to fit a "random wiggly curve" for each subject to the time-series of response latencies across the trials in the experiment. Within the linear mixed effect framework, the closest approximation would be a model including by-subject random intercepts and by-subject random slopes for `Trial`. But, as we shall see below, imposing linearity does not do justice to the data. The random factor smooths also take into account the "vertical positioning" of the wiggly curves over experimental time, i.e., they take care of what in the linear mixed effect model would be accounted for by means of random intercepts. For subjects, additional random slopes for `Frequency PC` and `Part-Whole Balance PC` were found to be also justified.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.7500	0.0728	-24.0328	< 0.0001
Word Length	0.0067	0.0040	1.6998	0.0892
SCC=TRUE	-0.0160	0.0168	-0.9518	0.3412
B. smooth terms	edf	Ref.df	F-value	p-value
smooth Frequency PC	2.3021	2.4940	46.2645	< 0.0001
smooth Part-Whole Balance PC : SCC=FALSE	1.4554	1.5540	12.6024	0.0001
smooth Part-Whole Balance PC : SCC=TRUE	1.0005	1.0006	17.4281	< 0.0001
random intercepts tone of first syllable	3.5901	5.0000	42.7653	< 0.0001
random intercepts tone of second syllable	0.1776	5.0000	0.0570	0.3675
random intercepts word category	0.8549	3.0000	4.6022	0.0822
smooth Positional Family Size PC : SCC=FALSE	1.0001	1.0001	0.5445	0.4606
smooth Positional Family Size PC : SCC=TRUE	1.0005	1.0007	9.6217	0.0019
random intercepts Word	367.1178	534.0000	2.3098	< 0.0001
random by-Subject slopes for Part-Whole Balance PC	25.6535	32.0000	9.1267	< 0.0001
random by-Subject slopes for Frequency PC	26.9802	32.0000	13.1872	< 0.0001
by-Subject random smooths for Trial	248.0165	296.0000	98.4261	< 0.0001

Table 7: Generalized Additive Model fitted to the negative inverse transformed lexical decision latencies of the smaller-scale multiple-subject study. SCC is a factor indicating membership of the strongly connected component of the compound graph.

In the main, the effects observed in the multi-subject experiment mirror those for the single-subject experiment. However, the effects of the tone of the second syllable, as well as that of word category, are lost, due to a lack of power. The effect of `Part-Whole Balance PC` and its interaction with membership in the strongly connected component of the compound graph was replicated. For words in the strongly connected component, the effect of `Part-Whole Balance PC` was somewhat reduced. An effect of `Positional Family Size PC` also re-emerged, but now its effect was strictly linear, with a negative slope. Figure 4 present the partial effects of these principal components, comparing the effects in Experiment 1 (upper panels) with those in Experiment 2 (lower panels).

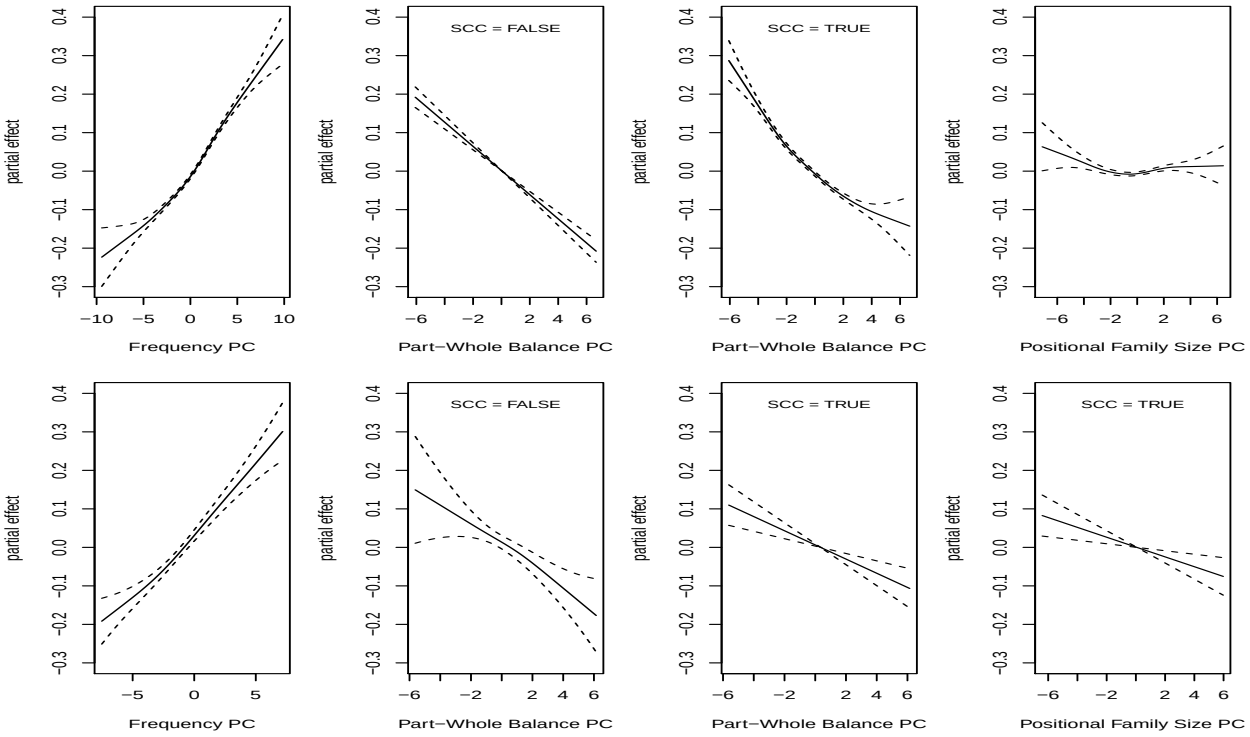

Figure 4: Smooths for the principal components for the single-subject data (top) and the multiple-subject data (bottom). SCC: factor denoting membership in the strongly connected component of the compound graph.

As for the single-subject experiment, we investigated the contributions of the predictors (or groups of predictors) in terms of the extent to which they contributed to reducing the AIC of the model. Table 8 indicates that subject and item variability dwarves the linguistic predictors. This pattern is strikingly different from that observed for the single-subject experiment, for which the first two principal components (`PC frequency` and `PC freq-fam`, in interaction with membership in the strongly connected component) effected the greatest changes in AIC. In other words, a design with multiple subjects comes at the cost of huge subject variablity, and huge variability with respect to how subjects respond to items.

By far the most important random-effect component in this model is given by the by-subject random smooths for `Trial`, visualized in Figure 5. As the experiment proceeded, subjects' performance fluctuated substantially, and non-linearly. Although for some subjects, these fluctuations were mild, other subjects showed performance that changed substantially. One subject started out as the slowest subject, but by the end of the experiment responded fastest, possibly indicating an effect of habituation to the task. Conversely, the subject starting out as the fastest responder became one of the slowest responders in the second half of the experiment. One subject revealed a highly oscillatory pattern, with tremendous slowing down, followed by speeding, up, in the last quarter of the experiment. We note here that the reduction in AIC afforded by the factor smooths, 7413.72, is substantially larger than the corresponding linear mixed-effects model with straight lines (obtained with random intercepts and random slopes) replacing the wiggly curves (6466.98).

An analysis of the subset of words with a second constituent in the strongly connected component was carried out to inspect whether the interaction of `Shortest Path Length` by `Frequency PC` by

models	AIC
+ Trial by Subject factor smooths	7413.72
+ Subject random intercepts and slopes	2176.57
+ Item random intercepts	954.61
+ Tones	2.96
+ Word Category	-0.99
+ Word Length	0.20
+ Frequency PC	1.47
+ Part-Whole Balance PC * SCC	3.55
+ Positional Family Size PC	2.03

Table 8: Reduction in AIC as predictors are added to an intercept only baseline model, for the multiple-subject data. SCC: factor denoting membership in the strongly connected component of the compound graph.

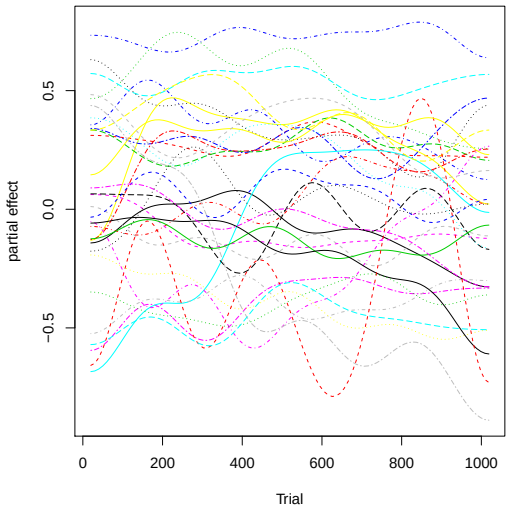

Figure 5: Factor smooths with shrinkage for Trial by Subject in Experiment 2. Each wiggly curve represents how a specific subject proceeds through the trials of the experiment. For instance, the initially fastest subject (light blue) ends the experiment with average speed, after having been one of the slower subjects in the second half of the experiment.

the second constituent being in use as a classifier would persist (model not shown). This interaction was again present, and as before, it was restricted to those compounds with a second constituent that is not in use as a classifier.

Finally, we note that the general inhibitory effect of Part-Whole Balance PC in Vietnamese replicated well in Experiment 2, providing further empirical support for the predictions of the ndl model. We therefore consider the learning model in some more detail.

Further modeling with naive discrimination learning

In the introduction, we observed that the ndl model predicted that Vietnamese lexemes are better learned when the corresponding two-syllabeme words are used more frequently, and are

learned less well the more the individual syllabemes are more entrenched in the sense that they are more frequently used, and used more often in other two-syllabeme words. This analysis shows that how well a lexeme is learned is itself co-determined by how its letter bigrams are used across the lexicon.

However, when reading a compound such as *tàu hoả*, the digraphs of the word will activate not only the lexeme of the compound (TRAIN), but also the lexemes of the constituent syllabemes (SHIP and FIRE). We therefore also calculated the model's support for the lexemes of the constituent syllabemes, expecting to find that greater support for the constituent syllabemes' lexemes gives rise to longer response latencies.[5] We therefore fitted a new GAMM to the response latencies of Experiment 1, with as predictors `Minutes, Session, Word Length`, membership in the strongly connected component (`SCC`), `Word Category, Tone, Compound Frequency`, and a tensor product smooth for the interaction of the NDL support for the lexemes of the compound and its syllabemes respectively. As some syllabemes occur only in compounds (compare *cran* in English *cranberry*), the analyses reported below are carried out on the 13681 compounds for which lexemes are available for the compound itself and for both its constituent syllabemes.

Compound frequency is incorporated in our analysis as an estimate of the a-priori probability that a word will be presented in the experiment. The greater the probability of correctly guessing what word will be shown on the screen, the faster a response can be initiated. (The compound frequency measure is theoretically well-motivated within the NDL learning framework, as relative frequencies can arise as a result of learning one-to-many mappings. The one-to-many mapping involved here is a subject's 'existence' as cue, and possible words in Vietnamese as outcomes.)

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.6621	0.0239	-69.4582	< 0.0001
Word Length	0.0190	0.0015	12.5854	< 0.0001
SCC = True	-0.0493	0.0054	-9.1795	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
smooth Log Compound Frequency	3.1282	3.9038	105.8299	< 0.0001
tensor product smooth for the three NDL measures	36.0092	46.2110	20.2480	< 0.0001
random effect Tone of First Syllabeme	3.1060	5.0000	2.8044	0.0013
random effect Tone of Second Syllabeme	0.0691	5.0000	0.0141	0.3889
random effect Word Category	6.3550	10.0000	6.2286	< 0.0001
smooth Minutes	4.1275	4.7890	39.9089	< 0.0001
smooth Session Number	8.4133	8.7888	38.0587	< 0.0001

Table 9: Generalized Additive Model fitted to the negative inverse transformed lexical decision latencies of the large single-subject study (edf: estimated degrees of freedom, SCC: factor denoting membership in the strongly connected component), with learning-based predictors.

The resulting model is summarized in Table 9. In what follows, we focus on the predictors of interest from a modeling perspective: The effect of a-priori probability, and the effects of the NDL support for the compound lexeme and its corresponding syllabemic lexemes. The upper left panel of Figure 6 presents the effect of compound probability, which is, as expected, facilitatory. The remaining panels visualize the three-way interaction of compound support by left and right syllabeme support. Each panel shows the fitted surface for a given pair of support measures, with other predictors in the model held constant at their most typical values. The upper right and lower left panels show that processing is delayed most for high syllabeme support and low compound support. Note, furthermore, that for the lowest values of syllabeme support, there is little effect of compound support. Finally, the lower right panel indicates that processing is optimal for syllabeme

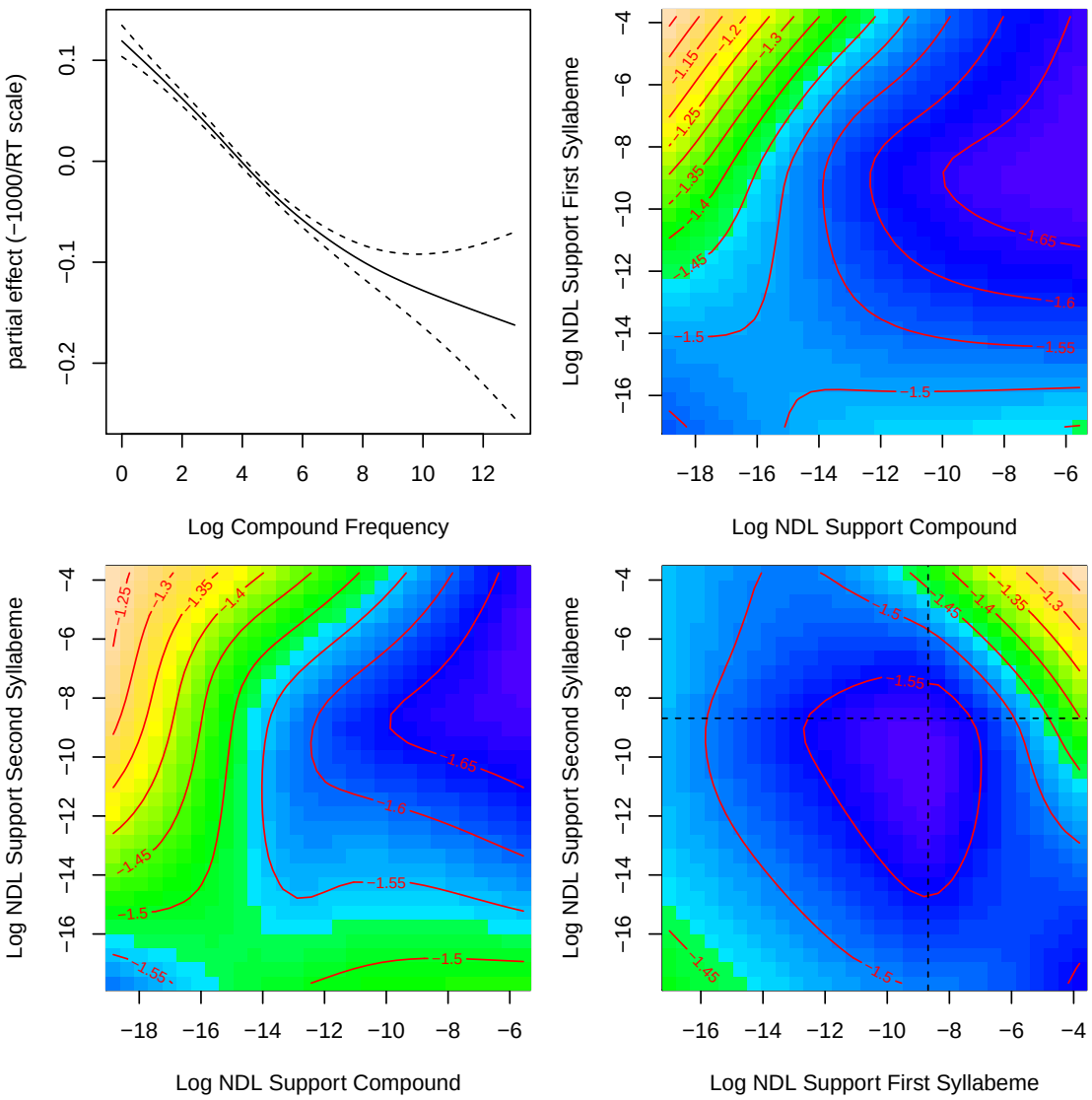

Figure 6: Partial effects of frequency and the interaction of the three NDL support measures in the GAMM fitted to the inverse-transformed response latencies of the single subject experiment. Darker shades of blue indicate shorter response latencies. Contour lines connect points with the same response latency. Values on the contour lines are on the -1000/RT scale.

support values close to the most typical values of syllabeme support (as indicated by the dashed lines, representing the medians, in the lower right panel). In other words, if a syllabeme has average (and hence well expected and least surprising) support, it is least intrusive in the visual lexical decision task.

It is noteworthy that this GAMM provides a much better fit to the data than the original model presented in Table 4. The model with NDL predictors has an AIC of 137.1. This compare very favorably to the model with the principal components replacing the NDL measures as predictors (AIC: 596.6) (Allowing for a four-way interaction of the three principal components and membership of the strongly connected component does not provide an improvement (AIC: 594.8) in goodness of fit with respect to the model with non-interacting principal components.)

In summary, we have shown that response latencies can be predicted with substantially greater accuracy when a learning approach is adopted. In this learning approach, there are two ways in which a compound's constituent syllabemes interfere and slow down comprehension.

The first kind of interference takes place during implicit learning, the never-ending process of adjusting the weights from orthographic cues to lexemic outcomes. Since compounds re-use syllabemes that often have their own meanings, and since these meanings are seldom contributing in a fully compositional way to the meaning of the compound (e.g., a 'fire engine' is, in English, an truck used to extinguish fires, whereas in Vietnamese, it is a vehicle, designed to drive on rail tracks, that used to be propelled by fire), when learning what a compound means, there is a constant tug of war between the cues and the compound lexeme on the one hand, and the cues and the syllabemic lexemes on the other.

To understand this tug of war, we have to take a step away from the intuitive (and behaviorist) idea of associative learning, according to which learning amounts to associations being formed in memory for co-occurring cues and outcomes. This intuitive view of learning ignores that *unlearning* takes place whenever cues fail to predict outcomes, a point emphasized by Rescorla (1988). Returning to the example from the introduction: Having whiskers is a cue to cats, rats, and rabbits. When whiskers are seen together with a rat, the weight on the link between whiskers and rat is strengthened, but at the same time, the weights on the links to cats and rabbits are *unlearned and weakened*, even though it is a fact about the world that cats and rabbits have whiskers (see also Marsolek, 2008, for unlearning in vision). This unlearning is one of the factors driving the inhibitory effect of `Part-Whole Balance PC` in the present experiments: The more frequent a constituent is and the less frequent the compound, the more the meaning of the compound will be unlearned from the cues of that constituent when that constituent is read in isolation (see also Ramscar et al., 2013, for more general consequences of unlearning).

The second kind of interference takes place during the event of compound reading itself: Intrusive, well-learned syllabemic lexemes become activated, just as *hat* in *that* is activated in English (Bowers et al., 2005; Baayen et al., 2007). To resolve the conflict between co-active lexemes, further control processes must be involved (see, e.g., Yeung et al., 2004; Ramscar and Gitcho, 2007). The greater the support for the intruding syllabemic lexemes, the more time is required by these control processes to resolve these conflicts.

As we did not obtain any evidence for an interaction involving the NDL measures and membership in the strongly connected component (SCC), it seems likely that the effect of SCC arises after the compound and syllabeme lexemes have been activated. Possibly, syllabemic lexemes in the strongly connected component of the compound graph generate, due to their higher interconnectedness, more predictions about lexemes they combine with. As these predictions do not match the visual input, the control processes have more evidence against such syllabemic lexemes, allowing faster responses (cf. the negative sign of the effect of SCC in Table 9).

A methodological note

When resources are limited, is it better to conduct a large study with one, or only a few, participants, or to conduct a study with more participants and fewer items?

The answer depends on the goal of one's study. If the goal is to study between-subject variation in language processing, obviously a multi-subject design is the appropriate choice. An important caveat here is that the majority of experiments in psychology and psycholinguistics make use of convenience samples of subjects — typically undergraduate, predominantly female, students of psychology (Francis et al., 2001; Sander and Sanders, 2006). Experiment 2 of the present study is no exception, with a majority of female participants, and with both males and females being university

students. As shown by (Kuperman and Van Dyke, 2011, 2013), substantial between-subject differences exist in reading skills (and reading habits) as a function of education and vocation. Thus, the multiple-subject experiment is revealing only about a very small, unrepresentative section of Vietnamese readers. Anyone interested in generalizing to a broader section of society should consider stratified random sampling from the full society.

The goal of the present study is not clarifying between-speaker differences in reading printed Vietnamese, but rather, exploring the consequences of experience with the lexical-distributional properties of Vietnamese for reading. A problem that arises here is that we do not have data on the experiences of individual subjects. All we have is an aggregate — the corpus — that cannot but be inaccurate for any individual reader.

Given the limitations of our current resources, the question is whether we learn more about the consequences of lexical-distributional predictors for lexical processing from a single-subject experiment, or from a multi-subject experiment with participants with a similar socio-economic background.

To address this question, we first assessed the adjusted R-squared obtained by fitting separate models with only lexical predictors for each of the 33 subjects in Experiment 2. We then compared the distribution of R-squared values with the corresponding distribution of R-squared values obtained by randomly sampling 500 data points (compounds) from Experiment 1, 30 times, and fitting the same model to these subsets of data. In the mean, the two distributions were indistinguishable, but the variance for the single-subject sample of R-squared values was significantly smaller ($p < 0.0001$, F-test). This is remarkable, as the subsamples cover a much wider range of words. It suggests that the between-subject variability in performance is much larger than the within-subject variability in performance.

This possibility receives further support when the amount of variance explained in the two experiments is scrutinized. The adjusted R-squared for Experiment 1 is 0.21, and that for Experiment 2, 0.59. However, most of the variance captured by Experiment 2 concerns between-subject variation. This becomes clear when we compare these adjusted R-squared values with those obtained by fitting models with all lexical predictors excluded, using only predictors such as `Trial`, `Minutes`, and `Session`. The adjusted R-squared for Experiment 1 is only 0.04, whereas for Experiment 2, it is 0.46. Thus, the bulk of the variance captured in our multi-subject experiment concerns subject variation. By contrast, the bulk of the variance for the single-subject experiment is captured by lexical-distributional predictors.

The advantage of having better coverage of the language with our single-subject experiment is illustrated by the interaction of the Frequency PC, the Shortest Path Length, and the use of the second constituent as Classifier. The fitted tensor surfaces for Experiments 1 and 2 are shown in the left and right panels of Figure 7 respectively. Due to data sparsity, the tensor for the multi-subject experiment (right panel) captures only the bottom half of the effect that emerges from the single-subject experiment (which has 30 times as many items). Thus, Experiment 1 emerges as more useful for understanding the linguistic aspects of lexical processing.

It might be argued that our single subject for Experiment 1 is, in some way, atypical. For instance, he might have been better motivated. On the other hand, at times, he might also have been more bored: The nonlinear pattern over sessions may well have been affected by a combination of the drudgery of performing yet another uninteresting lexical decision experiment and consideration of the number of sessions yet to be completed. Furthermore, our subject was an expatriate at the time of testing, which might have affected performance negatively. Fortunately, the very similar adjusted R-squared distributions for the 33 subjects of Experiment 2 and the 30 disjunct subsamples of 500 items from Experiment 1 suggest that the subject of Experiment 1 is not that different from other university-educated native speakers of Vietnamese sampled in Experiment 2.

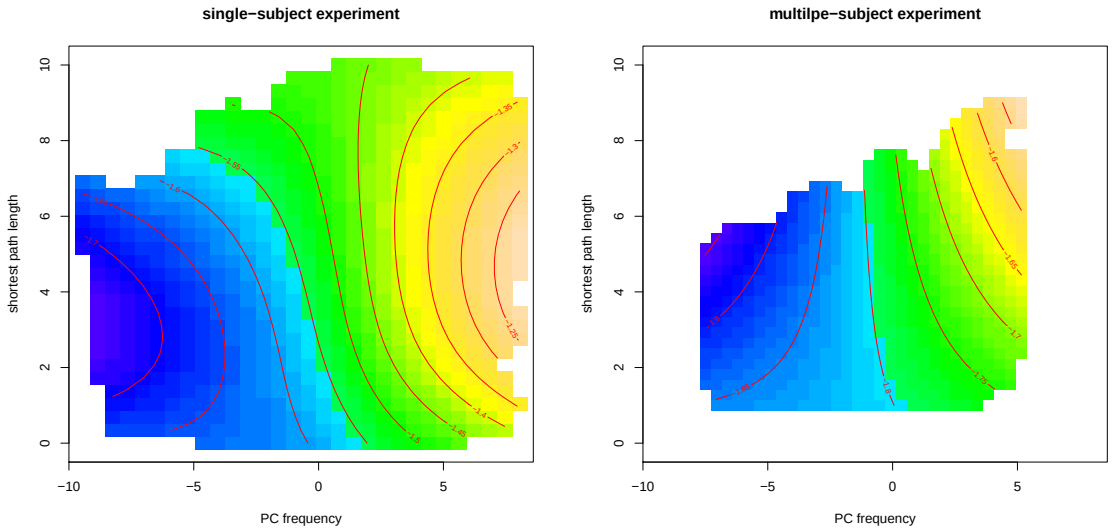

Figure 7: Tensor product surface for the interaction of Shortest Path Length and PC freq for compounds the second constituent of which is not in use as a classifier, in the single-subject (left) and multi-subject (right) experiment.

In the light of these considerations, we think that for languages with few speakers, or languages with few speakers with the necessary metalinguistic skills required for standard psycholinguistic behavioral paradigms, a comprehensive single-subject may therefore have advantages to offer when the focus of interest is on language rather than on socially-conditioned variation in language processing.

General Discussion

This study reports what is — to our knowledge — the first experimental study of Vietnamese. We have documented the effect on lexical decision latencies of a wide range of predictors, ranging from lexical tone to family size, and from membership in the strongly connected component to compound frequency.

One interesting result is the strong effect of lexical tone in the visual lexical decision task. The literature on the processing of tone is surprisingly limited. Cutler and Hsuan-Chih (1997) observed that same-different judgements were difficult for words differing only for tone, irrespective of whether subjects spoke Cantonese or not. Zhang and Damian (2009) observed faster responses in a tracking task for segments compared to tones. Zhao and Jurafsky (2009) reported for speech production a higher F_0 for lower-frequency words with mid tones. Shaw et al. (2014), also studying production, using electromagnetic articulography, observed independence of vowel and tonal targets. Nixon (2014) studied tone sandhi using the picture-word interference paradigm. The present study adds to this literature by demonstrating that the tones of Vietnamese come with specific processing costs, even in silent reading. For instance, in the multiple subject experiment, the *ngang* mid level tone elicited on average the shortest response latencies, while response latencies were longest for the *huyền* low falling (breathy) tone. In Vietnamese, tone is marked by diacritics on vowel letters. Nevertheless, these tiny diacritics are clearly highly discriminative, and give rise to strong effects in the reaction times. As these effects vary between whether the diacritic appears on the vowel in the first syllabeme or that on the second syllabeme (see Table 9), it is unlikely that the effect of the tone diacritics can be reduced to a purely orthographic effect. We think it is more likely that

it reflects the involvement of sound structure in reading (see also Carreiras et al., 2005; Lee, 2007; Winskel and Perea, 2014).

A second result of interest is the effect of membership in the strongly connected component of the directed compound graph of Vietnamese. Effects of network density have been reported for English (Baayen, 2010b), and the present results are encouraging enough to suggest it may be worth exploring whether these kind of "network effects" can be replicated in other languages as well.

The most surprising result that we obtained is that in Vietnamese, in contrast to English, constituent syllabemes interfere with reading. This interference was predicted by the NDL model, and was observed in two independent experiments. As NDL models for English correctly predict facilitation from constituents instead of inhibition, (see, e.g., Baayen et al., 2011, for one implementation, other implementations not shown here yield comparable results), the main source of this cross-language difference must reside in the distributional properties of the lexicons of English and Vietnamese. Our hypothesis is that Vietnamese, with its highly restricted syllable phonotactics, and orthographic conventions that space all compounds, is forced to overload its syllable-morphemes to a much higher extent to English. We think this stronger overloading lies at the heart of the Vietnamese constituent anti-frequency effect.

In the discriminative learning approach adopted in this study, there are no form units for syllabemes nor for compounds, and yet morphological effects are properly predicted. For a language traditionally described as isolating (recall the quote from Anderson cited in the introduction), this is an especially fitting result. Given the NDL model and the excellent predictivity of the predictors it offers, one might be tempted to conclude that Vietnamese compounds are 'just' two-syllable words, the syllables of which happen to have independent meanings, just as *hat* in *that* has its own meaning. This temptation should be resisted, however, as Vietnamese compounds show the same kind of weak, a-posteriori comprehensible compositionality that characterizes English compounds. In other words, we think that Vietnamese compounds are partially and idiosyncratically motivated, and hence motivated signs, albeit the product of long and equally idiosyncratic evolutionary paths through cultural history. Of course, this does not entail that a decomposition of Vietnamese compounds into constituent morphemic forms would play a role — to the contrary, no such form-driven decomposition takes place in our learning model. What does happen is that orthographic cues may co-activate the lexemes of constituent syllabemes, especially when these syllabemes have high frequencies of use compared to the compound itself. The resolution of the conflict between co-active syllabemic and compound lexemes may in turn have further repercussions at higher levels of cognition, leading to phenomena such as folk-etymologies and the intuitive feeling that the compounds in one's native language make eminent sense, which they do not (compare *tàu hoả* and *fire engine* in Vietnamese and English). The modeling of the consequences of these higher-level cognitive repercussions for lexical processing is a challenge for future research.

Was Anderson right in describing Vietnamese as a language "with nearly every word made up of one and only one formative"? Given the present results, the answer is both no and yes: No because compounds are rampant in Vietnamese, and yes, because compounds are more similar to two-syllable simple words than comparable compounds in English.

Notes

[1] We present the simulation first, and the experiments second, for expositional clarity. We note here that with respect to the "context of discovery", the experiments were run first. The anti-frequency effect observed in the reaction times then led us to test naive discrimination learning against the Vietnamese data.

[2] Baayen (2014) provides a short non-technical introduction to the GAMM. For examples of the use of generalized mixed-effects additive models in psycholinguistics, see Baayen (2014); Baayen et al. (2010); Tremblay and Baayen

(2010); Kryuchkova et al. (2012); DeCat et al. (2014b) and Balling and Baayen (2012), and for applications in linguistic studies, Wieling et al. (2011); Kösling et al. (2013); Wieling et al. (2014) and Tomaschek et al. (2013).

[3]Note that it is not necessary for a random-effect factor to have levels representing a sample of a much larger population. For such factors, just as for the present factors, the shrinkage estimates of the coefficients afford more precise estimates for when the same levels are sampled in a future replication study. When the population is large, as typically is the case for subjects and items, then the mixed model provides an estimate for unknown subjects and items, thanks to the fixed-effect estimates for the population. For random effect factors such as Tone and Word Category, we have no interest in unsampled tones or word categories, as there are none. Nevertheless, we can profit from the shrinkage estimates to protect against overfitting with many factor levels while bringing systematic non-independence related to Tone and Word Category into the model.

[4]Akaike's information criterion, or AIC (Akaike, 1974) is an information-theoretic measure of goodness of fit. Smaller values indicate a better fit.

[5] Modeling with NDL requires decisions about what form information to use for cues, and what lexemic information to use for the outcomes. With respect to the cues, we explored letter pairs and letter trigrams. With respect to the outcomes, we compared models using as outcomes the lexemes of the compound together with the lexemes of its constituents with models using as outcomes only the compound lexeme. The latter models outperformed the former when pitted against reaction times. We therefore report results only for the best model, using letter bigrams as cues, and non-decompositional lexemic representations as outcomes.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Anderson, J. R. (1985). Typological distinctions in word formation. In Shopen, T., editor, *Language typology and syntactic description*, pages 3–56. Cambridge University Press, Cambridge.

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes*. The MIT Press, Cambridge, Mass.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press, Cambridge, U.K.

Baayen, R. H. (2010a). Assessing the processing consequences of segment reduction in dutch with naive discriminative learning. *Lingue & Linguaggio*, 9:95–112.

Baayen, R. H. (2010b). The directed compound graph of English. an exploration of lexical connectivity and its processing consequences. In Olsen, S., editor, *New impulses in word-formation (Linguistische Berichte Sonderheft 17)*, pages 383–402. Buske, Hamburg.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, 11:295–328.

Baayen, R. H. (2014). Multivariate Statistics. In Podesva, R. J. and Sharma, D., editors, *Research Methods in Linguistics*. Cambridge University Press, Cambridge.

Baayen, R. H., Hendrix, P., and Ramscar, M. (2013). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Language and Speech*, 56:329–347.

Baayen, R. H., Kuperman, V., and Bertram, R. (2010). Frequency effects in compound processing. In Scalise, S. and Vogel, I., editors, *Compounding*. Benjamins, Amsterdam/Philadelphia.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118:438–482.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H., Wurm, L. H., and Aycock, J. (2007). Lexical dynamics for low-frequency complex words. a regression study across tasks and modalities. *The Mental Lexicon*, 2:419–463.

Balling, L. and Baayen, R. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, 125:80–106.

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., and Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology:General*, 133:283–316.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics. Identifying Influential Data and sources of Collinearity.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Blevins, J. P. (2003). Stems and paradigms. *Language*, 79:737–767.

Bowers, J., Davis, C., and Hanley, D. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language*, 52:131–143.

Carreiras, M., Ferrand, L., Grainger, J., and Perea, M. (2005). Sequential effects of phonological priming in visual word recognition. *Psychological Science*, 16(8):585–589.

Chen, Q. and Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological review*, 119(2):417.

Colé, P., Segui, J., and Taft, M. (1997). Words and morphemes as units for lexical access. *Journal of Memory and Language*, 37(3):312–330.

Cutler, A. and Hsuan-Chih, C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics*, 59(2):165–179.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, 47(2):109–121.

DeCat, C., Baayen, H., and Klepousniotou, E. (2014a). Electrophysiological correlates of noun-noun compound processing by non-native speakers of English. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*, pages 41–52, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

DeCat, C., Klepousniotou, E., and Baayen, R. H. (2014b). Representational deficit or processing effect? A neuro-psychological study of noun-noun compound processing by very advanced L2 speakers of English. *Frontiers in Psychology (Language Sciences)*, page under revision.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42(2):488–496.

Forster, K. I. and Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, and Computers*, 35(1):116–124.

Francis, B., Robson, J., and Read, B. (2001). An analysis of undergraduate writing styles in the context of gender and achievement. *Studies in Higher Education*, 26(3):313–326.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall, London.

Hoàng, P., editor (2000). *Từ điển tiếng Việt [Vietnamese Dictionary]*. Khoa học Xã hội, Hà Nội. Viện Ngôn ngữ học.

Keuleers, E. and Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behaviour Research Methods*, 42(3):627–633.

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44(1):287–304.

Kryuchkova, T., Tucker, B. V., Wurm, L., and Baayen, R. H. (2012). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. *Brain and Language*, 122:81–91.

Kuperman, V., Bertram, R., and Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, 23:1089–1132.

Kuperman, V., Schreuder, R., Bertram, R., and Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP*, 35:876–895.

Kuperman, V. and Van Dyke, J. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, 65:42–73.

Kuperman, V. and Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, 39(3):802.

Kösling, K., Kunter, G., Baayen, H., and Plag, I. (2013). Prominence in triconstituent compounds: Pitch contours and linguistic theory. *Language and Speech*.

Lee, C.-Y. (2007). Does horse activate mother? Processing lexical tone in form priming. *Language and Speech*, 50(1):101–123.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge University Press, Cambridge.

Marsolek, C. J. (2008). What antipriming reveals about priming. *Trends in Cognitive Science*, 12(5):176–181.

Matthews, P. H. (1974). *Morphology. An Introduction to the Theory of Word Structure*. Cambridge University Press, London.

Milin, P., Ramscar, M., Choc, K., Baayen, R. H., and Feldman, L. B. (2014). Processing partially and exhaustively decomposable words: An amorphous approach based on discriminative learning. *Manuscript, University of Novi Sad*.

Miwa, K., Libben, G., Dijkstra, T., and Baayen, H. (2014). The time-course of lexical activation in Japanese morphographic word recognition: Evidence for a character-driven processing model. *The Quarterly Journal of Experimental Psychology*, 67(1):79–113.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., and Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1271–1278.

Mulder, K., Dijkstra, T., Schreuder, R., and Baayen, R. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language*, 72:59–84.

Nguyễn, D. H. (1997). *Vietnamese: Tiếng Việt không son phấn*. John Benjamins, Amsterdam.

Nguyễn, T. G. (1996). *Từ và nhận diện từ tiếng Việt [Words and recognizing words in Vietnamese]*. Giáo Dục, Hà Nội.

Nguyễn, T. G. (2011). *Vấn đề "từ" trong tiếng Việt [The issues of "word" in Vietnamese]*. Nhà xuất bản Giáo Dục, Hà Nội.

Ngô, T. N. (1984). The syllabeme and patterns of word formation in Vietnamese. New York University dissertation.

Nixon, J. S. (2014). Sound of mind. behavioural and electrophysiological evidence for the role of context, variation, and informativity in human speech processing.

Pham, H. (2014). *Visual processing of Vietnamese compound words: A multivariate analysis using corpus linguistic and psycholinguistic paradigms*. PhD dissertation, University of Alberta, Edmonton.

R Core Team (2014). R: A language and environment for statistical computing.

Ramscar, M. and Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends In Cognitive Science*, 11(7):274–279.

Ramscar, M., Hendrix, P., Love, B., and Baayen, R. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, 8:450–481.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957.

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *American Psychologist*, 43(3):151–160.

Sander, P. and Sanders, L. (2006). Rogue males: Sex differences in psychology students. *Eletronic Journal of Research in Educational Psychology*, 8(4):1.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27.

Shaoul, C., Arppe, A., Hendrix, P., Milin, P., and Baayen, R. H. (2013). *ndl: Naive Discriminative Learning*. R package version 0.2.14.

Shaw, J., Chen, W., Proctor, M. I., Derrick, D., and Dakhoul, E. (2014). On the inter-dependence of tonal and vocalic production goals in Chinese. In *Proceedings of the 10th ISSP, Cologne, May 5–8.*, pages 395–398, Cologne. ISSP.

Tomaschek, F., Wieling, M., Arnold, D., and Baayen, R. H. (2013). Frequency effects on the articulation of German i and u: evidence from articulography. In *Proceedings of Interspeech, Lyon*, pages 1302–1306.

Tremblay, A. and Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In Wood, D., editor, *Perspectives on Formulaic Language: Acquisition and communication*, pages 151–173. The Continuum International Publishing Group, London.

Wagner, A. and Rescorla, R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical Conditioning II*, pages 64–99. Appleton-Century-Crofts, New York.

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language*, 90(3):669–692.

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6(9):e23613.

Winskel, H. and Perea, M. (2014). Does tonal information affect the early stages of visual-word processing in Thai? *The Quarterly Journal of Experimental Psychology*, 67(2):209–219.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73:3–36.

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall/CRC, New York.

Yeung, N., Botvinick, M. M., and Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, 111(4):931.

Zhang, Q. and Damian, M. F. (2009). The time course of segment and tone encoding in Chinese spoken production: an event-related potential study. *Neuroscience*, 163(1):252–265.

Zhao, Y. and Jurafsky, D. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics*, 37(2):231–247.