Article

Language & Linguistics 語言暨語言學

Language and Linguistics
16(6) 791–818

DOI: 10.1177/1606822X15602606
lin.sagepub.com

SAGE

Markedness and Lexical Typicality in Mandarin Acceptability Judgments*

James Myers

National Chung Cheng University

It has long been known that native speakers judge nonlexical forms as more acceptable the more lexically typical they are (i.e. similar to real words). It has also been shown that speakers judge less marked (i.e. universally more natural) structures as more acceptable than more marked structures. In this study, we investigate for the first time how markedness and lexical typicality interact, using data from a large corpus of experimentally collected Mandarin native-speaker judgments of nonlexical syllables. We defined the lexical typicality of a test item in terms of how many lexical Mandarin syllables share the item's onset consonant, and defined markedness in terms of how many phoneme inventories have this consonant cross-linguistically. Consistent with prior research, both markedness and lexical typicality improved acceptability, but the two factors also interacted positively: the lexical typicality effect was stronger for less marked forms than for more marked forms. The same interaction appears in a reanalysis of English nonword judgment data. This interaction is not predicted by standard Optimality Theory, but it can be formalized with conjunctive coordination, the inverse of the more familiar local conjunction, whereby a coordinated constraint is obeyed if and only if both of its component constraints are obeyed.

Key words: acceptability experiments, English, Mandarin, markedness, Optimality Theory

1. Introduction

Generative linguists are interested in two key questions: what productive knowledge native speakers have about their language, and how this knowledge relates to the universal human language faculty. In lexical phonology, including phonotactics, productive knowledge involves a speaker's sense of what makes a phonological form typical for the speaker's lexicon. English speakers, for example, know that *blick*, though not an actual English word, is a possible word, while *bnick* never could be (Chomsky & Halle 1965). Simplifying, this is because /bl/ appears in real English words like *black* while /bn/ appears in no English words; that is, *blick* is more lexically typical than *bnick*. Meanwhile, phonologists are also obliged to search for universal principles that can explain

* Research for this paper was supported by National Science Council (Taiwan) grants NSC 100-2410-H-194-109-MY3 and NSC 101-2410-H-194-115-MY3, codirected with Jane Tsay. Many thanks to the experimental participants, research assistants (especially Chia-Wen Lo and Chen-Tsung Yang), two anonymous reviewers, and audiences at the 45th Annual Meeting of the Societas Linguistica Europaea in Stockholm, the Workshop on Phonological Markedness at National Tsinghua University, Hsinchu, Taiwan, and the 4th International Theoretical Phonology Conference at National Chengchi University, Taipei. Despite its long gestation, this paper is just part of a continuing work in progress, so reader suggestions are still most welcome!

such language-specific patterns. For example, /bn/ may be more marked (unnatural) than /bl/ because consonant clusters tend to favor larger sonority contrasts, and /b/ is closer to /n/ than to /l/ in sonority (e.g. Selkirk 1982).

Both lexical typicality and markedness have been shown to affect phonological processing, including the still-mysterious processes involved in making acceptability judgments. Consistent with Chomsky & Halle's (1965) intuitions, experiments (starting with Greenberg & Jenkins 1964) have consistently shown that nonwords with greater lexical typicality are judged as more acceptable. It is becoming increasingly well-established that phonological processing is also easier, and acceptability higher, for less marked nonwords, even after controlling for lexical typicality (see review in Hayes et al. 2009, but see Moreton & Pater 2012 for evidence that phonetic markedness has only weak and unstable effects on learning novel grammars).

In Optimality Theory (OT; Prince & Smolensky 2004), markedness is expressed with universal constraints, and lexical typicality is expressed as language-specific rankings. Both types of experimental findings can thus be readily accommodated in an OT framework, as we review in §2. Roughly speaking, the more a nonword's markedness constraint violation profile matches that of real lexical items, or the fewer markedness constraints it violates overall, the more acceptable it is.

As we also show, OT predicts that markedness and lexical typicality should have independent effects on acceptability judgments. However, there are good reasons to expect that markedness and lexical typicality should actually interact. Specifically, markedness (in its positive sense, i.e. naturalness) should enhance the effect of lexical typicality: speakers should be more sensitive to how lexically typical a form is, if that form is also less marked. This expectation starts from the assumption that the markedness of a form reflects the universal forces involved in its learning and processing. This means that it should be more difficult for speakers to develop and maintain well-defined mental representations for marked forms. This, in turn, should make it more difficult for them to sense the lexical typicality of marked forms relative to a specific lexicon.

In §3, we confirm this expectation in an acceptability judgment experiment on Mandarin involving over 3,000 nonlexical test items and over 100 native speakers. Since our claim goes beyond any one language, however, in §4 we also test it in a reanalysis of the English judgment data of Hayes & White (2013); again we observe an enhancing interaction between markedness and lexical typicality. In §5 we show that one way to formalize this interaction within OT is to adopt a hitherto underused variant of local constraint conjunction (see Smolensky 1993, 2006) called conjunctive coordination (Crowhurst & Hewitt 1997). In §6 we provide brief concluding remarks.

2. Lexical typicality, markedness, and Optimality Theory

Although most phonologists continue to focus their analyses on lexically attested forms, it is generally understood that the aims of generative phonology require demonstrating the productivity of phonological knowledge, something that can only be done by testing nonlexical items (just as syntacticians routinely test the acceptability of novel sentences). Fortunately, it is becoming more common for phonologists to test grammatical hypotheses by experimentally collecting acceptability judgments or other responses that require phonological processing (see Kawahara 2011 for a review). In this section we briefly review two types of such experiments: those that test the effects of lexical

typicality (§2.1) and those that test the effects of markedness (§2.2). We then show that standard OT predicts that these two factors should not interact (§2.3). Finally, we argue that there are reasons to doubt this prediction, in light of broader principles of cognitive psychology (§2.4).

2.1 Lexical typicality

Numerous experiments have shown that phonological acceptability judgments are improved by lexical typicality, that is, the similarity of nonword test items to lexical forms. The positive correlation of gradient acceptability with lexical typicality has been experimentally demonstrated for many languages, including English (e.g. Albright 2009; Bailey & Hahn 2001; Coetzee 2008; Coleman & Pierrehumbert 1997; Greenberg & Jenkins 1964; Hayes & Wilson 2008; Ohala & Ohala 1986; Vitevitch et al. 1997), Turkish (Zimmer 1969), Arabic (Frisch & Zawaydeh 2001), Hebrew (Berent & Shimron 1997), Tagalog (Zuraw 2007), Japanese (Kawahara & Kao 2012), Cantonese (Kirby & Yu 2007), and Mandarin (Myers & Tsay 2005; Wang 1998).

The precise nature of lexicality typicality, that is, the metric defining similarity between nonword test items and actual lexical items, is not yet clear. The literature often discusses two types, the more holistic metric of neighborhood density (the number of lexical items that differ from the test item in exactly one segment; see Bailey & Hahn 2001 for a review and alternative quantifications), and the more analytical metric of phonotactic probability (the lexical type frequency of the test item's segments or segment substrings; again, see Bailey & Hahn 2001 for complexities). Evidence suggests that these metrics play distinct roles in phonological processing (e.g. Stockall et al. 2004; Vitevitch & Luce 1999), including in acceptability judgments (e.g. Bailey & Hahn 2001).

Lexical typicality involves phonologically more sophisticated metrics as well. Some experiments suggest that quantifying the similarity between test items and lexical items crucially depends on phonological features, not just whole segments (Albright 2009; Hahn & Bailey 2005), while others have reported that acceptability judgments are affected by formal phonological constraints like the Obligatory Contour Principle (OCP; Leben 1973; McCarthy 1979), even when neighborhood density and phonotactic probability are controlled (Coetzee 2008; Frisch & Zawaydeh 2001). Thus lexical typicality is at least partially encoded in terms that theoretical phonologists, not merely psycholinguists, can appreciate.

The correlation between acceptability and lexical typicality is also entirely consistent with OT (e.g. Boersma & Hayes 2001; Coetzee 2008). Fundamentally this is because an OT constraint ranking for a specific lexicon is derived from that lexicon, whether by hand or via a formal learning algorithm. At the very least, an OT ranking will distinguish between attested forms and systematic gaps, and of course attested forms will necessarily be more lexically typical than systematic gaps. Thus even though the oldest OT learning algorithm, the Error-Driven Constraint Demotion algorithm of Tesar & Smolensky (1998), was not designed to capture gradient acceptability judgments, it does capture lexical typicality in this basic sense. Other learning algorithms are explicitly intended to yield gradient scores, and these do indeed correlate well with gradient acceptability judgments, as has been shown, for example, for the Gradual Learning Algorithm of Boersma & Hayes (2001).

The reason why OT grammars are capable of generating gradient scores that express lexical typicality, and thus help predict acceptability, is that they are closely related to the statistical method of regression modeling, in which a best-fit line is found for a scatterplot of data points. This

is because OT is derived historically from Harmonic Grammar (HG; Legendre et al. 1990), in which constraints are weighted rather than ranked, with the final score, technically called harmony, derived by summing the weighted constraint values for each candidate's violation profile. For any given lexicon, HG yields a regression equation, where the dependent variable is harmony, the independent variables are the constraints, and the coefficients (slopes) are the weights. When implemented via maximum entropy, as it most commonly is (e.g. Hayes & Wilson 2008), learning in HG is mathematically equivalent to logistic regression (Berger et al. 1996; Malouf 2002), familiar to linguists as the heart of the VARBRUL sociolinguistic analysis tool (Mendoza-Denton et al. 2003). Regardless of the algorithm (see Potts et al. 2010 for an alternative), the more frequently an HG learner encounters a phonological pattern in the lexicon, the greater the weight for any constraint that favors this pattern, and therefore the better the HG grammar reflects lexical typicality.

OT is a special case of HG where the weights are restricted such that for any given constraint weight *w*, the absolute value of the summed weights times maximum violations for all lower-ranked constraints is less than that for *w* (Prince & Smolensky 2004). This restriction makes it impossible for lower-ranked constraints in OT to 'gang up' and override higher-ranked constraints. These considerations imply that even hand-derived OT grammars, or the overtly categorical OT grammars learned via the algorithm of Tesar & Smolenksy (1998), actually evaluate forms on a quantitative, gradient scale, a conclusion that we will exploit below. In any case, the close connection between OT and best-fit regression lines shows why it is inevitable that OT grammars help predict acceptability judgments.

To make this abstract discussion a bit more concrete, consider the OT analysis that Coetzee (2008) gives for his English acceptability judgment experiment. Coetzee starts by noting that the English lexicon has many forms like *state*, where an initial /s/ is followed by /t/, vowel, and another /t/, but no morphemes like **spape* or **skake* (see Davis 1989 for an earlier analysis of this pattern). Coetzee goes on to observe (p. 229) that the *skVk constraint must actually be somewhat weaker than the *spVp constraint, in that only the former can be violated by slight deviations from the template. Thus English has words like *skag* (ending in a voiced velar rather than /k/), and *skunk* and *skulk* (with an intervening consonant), but no exceptions involving labials: **spab*, **spump*, **spulp*. These observations lead him to posit a family of *sCVC constraints, ranked as in (1), where the hierarchy of markedness constraints is split by faithfulness constraints, here symbolized with FAITH, that enforce the lexical preservation of certain universally marked forms (as *state* presumably is, since it violates *stVt). Note that morphemes that exactly match the templates /spVp/ and /skVk/ are banned from the lexicon, given their ranking above FAITH, but the relative weakness of other velar constraints like *skVg and *skVCk suggests that *skVk is actually outranked by *spVp.

(1) *spVp » *skVk » FAITH » *stVt

In other words, the ranking in (1) is ultimately motivated by lexical typicality: forms like *state*, or even nonwords like *stote*, better reflect the English lexicon than forms like *skake*, which in turn are more lexically typical than forms like *spape*. Coetzee then goes on to show that the ranking in (1) correlates with the ranking of the associated acceptability judgments. Since the test items were controlled for neighborhood density and phonotactic probability, Coetzee concludes that the *sCVC constraint family, ranked as in (1), plays an active role in describing the mental grammars of native English speakers.

We illustrate Coetzee's results schematically in (2), using examples of his test items (though his actual design involved test pairs). The symbol ☞ represents acceptance by native speakers (rather than grammaticality per se), with ☞☞ used to represent strong acceptance, to express that acceptability is inherently gradient. This is not an OT tableau, since the surface forms being evaluated are not competing candidates; the point is merely to compare their relative harmony. Note also that the FAITH constraint is not violated by any of the items; these are nonwords, so they have no underlying forms to be faithful to.

(2)

	*spVp	*skVk	FAITH	*stVt
[spɪp]	*			
☞ [skæk]		*		
☞☞ [stɔ:t]				*

To capture formally the correlation between gradient acceptability and lexical typicality, we can assign HG-style weights to the OT constraints in accordance with the restriction noted above. As shown in (3), this can be achieved by using an exponential function for the weights (Keller 2006), here powers of two (assuming no constraint is ever violated more than once by any given form). This means setting the lowest weight to 1 (= 2^0), the next-lowest weight to 2 (= 2^1), the third-lowest to 4 (= 2^2), and so on. If we code each violation as –1 and each nonviolation as 0, we can then compute the harmony for each form simply by multiplying each weight by each violation code and summing them up. For example, [skæk] gets a score of –4 (= 8·0 + 4·(–1) + 2·0 + 1·0). The result is a set of harmony scores that correlate with acceptability (higher scores, i.e. negative scores closer to zero, indicate greater acceptability).

(3)

		Weights:	8	4	2	1
			*spVp	*skVk	FAITH	*stVt
	–8	[spɪp]	*			
☞	–4	[skæk]		*		
☞☞	–1	[stɔ:t]				*

2.2 Markedness

The advantage of OT constraints in describing acceptability is not merely that they can capture more of the judgment data than neighborhood density and phonotactic probability alone. More importantly, OT constraints are also assumed to describe markedness, universal bits of phonological knowledge presupposed in a speaker's knowledge of any specific language. Thus Coetzee (2008) has to take the trouble to argue that his arcane-seeming *sCVC constraint family is actually derived from the OCP (e.g. *pVp) and other universal constraints against falling sonority in onset clusters

(e.g. *sp), combined through local constraint conjunction (Smolensky 1993, 2006; we return to this device later in the paper).

OT leads us to expect that acceptability judgments and other phonological processes should also be sensitive to markedness per se, even if lexical typicality is controlled. Experiments that test this prediction have been called 'UG experiments' (Hayes et al. 2009), since they address Universal Grammar rather than language-specific grammars. UG experiments have been run on speakers of English (Berent et al. 2007; Hayes & White 2013), Hungarian (Hayes et al. 2009), and Hebrew (Berent et al. 2012).

Consider the UG experiment of Hayes & White (2013). They began by using the HG phonotactic learner of Hayes & Wilson (2008) to model the English lexicon. This learner generates phonotactic constraints by systematically combining features and then fitting the constraint weights to lexical typicality. Since the constraint set is not fixed beforehand, the algorithm tends to generate both universal markedness constraints and bizarre language-specific constraints. Examples of each type are illustrated in (4).

(4) a. Natural: $*[-\text{SYL},+\text{HI}]]_{\text{CODA}}$ (No coda glides)
b. Unnatural: $*_{\text{WORD}}[[-\text{DIPHTHONG},+\text{RND},+\text{HI}]$ (No word-initial /u, ʊ/)

In their experiment, each constraint-violating test item was paired with another nonword that matched it as closely as possible, except that it obeyed the relevant constraint. The item pairs used to test the constraints in (4a) and (4b) are shown in (5a) and (5b), respectively, with the violating item shown first for each pair (the experimental participants were also shown the italicized spellings). Note that *jouy* [dʒɑʊj] violates the constraint in (4a) banning coda glides, here /j/, because the /ʊ/ is assumed to be part of the nuclear diphthong /ɑʊ/.

(5) a. *jouy* [dʒɑʊj] versus *jout* [dʒɑʊt]
tighw [taɪw] versus *tibe* [taɪb]
b. *ooker* [ˈʊkəˌ] versus *ocker* [ˈɑkəˌ]
utrum [ˈutɹəm] versus *otrum* [ˈoʊtɹəm]

English native speakers then judged the acceptability of each item, separately, on an open-ended continuous scale (using the magnitude estimation technique; Bard et al. 1996). As one might expect given the above examples, items that violated natural constraints like (4a) were judged significantly worse than those that violated unnatural constraints like (4b), in comparison to their nonviolating controls. This was so even though the HG learner derived both types of constraints equally readily from the English lexicon. Thus it appears that speakers have some sense of what makes a constraint natural, a sense that cannot have come from lexical experience alone.

Such markedness effects can be understood in terms of the Emergence of the Unmarked (TETU; Becker & Potts 2011; McCarthy & Prince 1994), whereby the effects of lower-ranked constraints emerge when higher-ranked constraints do not discriminate among competing candidates. In this case, the higher-ranked constraints are the faithfulness constraints that potentially allow exceptions to certain universal markedness constraints in English. Given that the test items are all nonwords, FAITH does not apply, allowing the markedness constraints to emerge.

However, markedness constraints can only emerge if they actually exist, or at least are ranked highly enough (Hayes et al. 2009 show that even unnatural constraints have *some* effect on acceptability). This is presumably the case for the natural constraint in (4a), but not for the unnatural constraint in (4b). We illustrate this contrast in (6) using the same conventions as in our earlier discussion of Coetzee (2008): the total number of violations of genuine markedness constraints is greater in (6a) (one) than in (6b) (zero).

(6) a. Natural (existing) markedness constraint: acceptability difference

		Weights:	2	1
			FAITH	*[–syl,+hi]]$_{Coda}$
	–1	[dʒaʊj]		*
☞	0	[dʒaʊt]		

b. Unnatural (nonexistent) markedness constraint: no acceptability difference

		Weights:	2	1
			FAITH	*$_{Word}$[[~~–DIPHTHONG, +RND, +HI~~]
☞	0	[ʊkə˞]		
☞	0	[ɑkə˞]		

2.3 Markedness and lexical typicality interactions in standard OT

OT thus seems capable of handling both lexical typicality effects and markedness effects in acceptability judgments. As we have seen, demonstrations of the latter effects have involved experimental designs that control for the former effects. However, this kind of design crucially assumes that the two factors do not interact. If lexical typicality effects actually vary in strength depending on the degree of markedness, then simply controlling one or the other will miss this crucial fact.

As it turns out, standard OT predicts that lexical typicality and markedness do not interact. To appreciate this, it is important to recall that in OT, phonological forms can be encoded solely in terms of their constraint violation profiles (Golston 1996). For example, a form violates the constraint *[+F] if and only if it contains the feature value [+F], so the descriptions on either side of 'if and only if' are equivalent (distinctive features themselves have been shown to be derivable via the ranking of phonetically detailed constraints; Kirchner 1997). Thus the more marked items in a lexicon are not represented qualitatively differently from the less marked items; they simply violate more markedness constraints. Since a grammar in standard OT is defined solely by the ranking of universal constraints, the lexically typical items in a given language are those that violate the markedness constraints ranked low in the language, consistent with the grammar. Atypical items instead obey these constraints and/or violate markedness constraints that are usually obeyed in the language.

With this as background, now imagine three universal markedness constraints, A, B, C, yielding forms like [ABC] (i.e. a form that obeys all three constraints), [aBC] (violating only constraint A), or [abC] (more marked than [aBC], since it violates two markedness constraints instead of just

one). Now imagine that we want to test for the effect of lexical typicality on acceptability in a language with the ranking A » B » FAITH » C (similar to the partial grammar discussed by Coetzee 2008). The test items are all nonlexical, so the constraint FAITH is irrelevant. In (7) we see two sets of test items from our experiment, where the items in (7a) are all less marked (one constraint violation each) and those in (7b) are all more marked (two constraint violations each). Weights are posited and harmony is computed in the usual way. Note that harmony, which should correlate with acceptability, reflects both markedness (the total number of constraint violations) and lexical typicality (the language-specific rankings/weightings of the violated constraints).

Now we ask whether OT predicts that the participants in our imaginary experiment should more readily distinguish among the less marked items in (7a) or among the more marked items in (7b). We can see that the ranges of the two sets of harmony scores are the same, namely 3 (–1 – (–4) in (7a) and –3 – (–6) in (7b)). To incorporate all data points into our measure of variability, we can compute the standard deviation (*SD*), which represents the average distance of each data point from the average of the whole data set (for details, see any introductory statistics textbook). This gives us *SD* = 1.53 for both (7a) and (7b). In other words, OT predicts exactly the same amount of variation in harmony across the less marked items as across the more marked: lexical typicality and markedness do not interact.

(7) a. Less marked items

Weights:		4	2	1
		A	B	C
–4	aBC	*		
–2	AbC		*	
–1	ABc			*

b. More marked items

Weights:		4	2	1
		A	B	C
–6	abC	*	*	
–5	aBc	*		*
–3	Abc		*	*

This result follows directly from the mathematical basis of OT. As a special case of HG, an OT grammar represents a regression equation like that in (8). Readers may recognize this as a linear equation (i.e. with only one constraint it would literally plot a straight line).

(8) $\text{Harmony} = \text{Weight}_A \times A + \text{Weight}_B \times B + \text{Weight}_C \times C$

A key property of a linear equation is that it is additive, which means that we can learn all we need to know about the effect of one variable independently of all of the other variables. This

property is of great practical importance to phonological analysis, since it means that we can study any single constraint without testing it under all possible values of every other constraint in the grammar. In other words, the constraints in standard OT do not interact with each other. The overall harmony will increase when each additional constraint is violated, but always by a fixed amount determined solely by that constraint's weight; the other constraints do not modulate this amount. Therefore, markedness, which reflects the total number of markedness constraints that are violated by a given form, and lexical typicality, which reflects the constraint weighting, do not interact either.

2.4 Why markedness and lexical typicality should interact

Despite the predictions of OT, there are very good reasons to expect that markedness does indeed modulate the effect of lexical typicality, specifically by enhancing it in less marked forms. This claim will later be shown to be consistent with a particular formal OT device, but here we motivate it in a theory-neutral way.

As already mentioned in the introduction, the central insight is that the effect of lexical typicality on acceptability depends on the ability of judges to discriminate among forms, and it should be harder to discriminate among items that are intrinsically difficult to process than among items that are easier to process. OT grammarians already assume that markedness acts as a filter on learning, and UG experiments suggest that it acts as a filter on adult processing as well. Therefore, speakers should not be able to sense lexical typicality equally efficiently in both marked and unmarked forms, contrary to what standard OT seems to imply.

The principle underlying this conclusion applies far more generally than phonology, or even language. A case in point is the well-established observation in the cognitive psychology literature that upside down faces are far harder to process (e.g. identify or discriminate) than upright faces. A particularly striking instance of this is the so-called Thatcher illusion (first reported in Thompson 1980, when Margaret Thatcher was Prime Minister of the United Kingdom), in which a photograph of Thatcher's face was distorted by flipping the eyes and mouth upside down. When upright, the 'thatcherized' face is immediately distinguishable from the original photograph, but when upside down, the distorted and normal faces are very hard to discriminate (Bartlett & Searcy 1993).

What makes the Thatcher illusion relevant here is that the distorted faces can only be identified when we are allowed to use our highly efficient face-processing system, which only turns on for upright faces. Just as in language acquisition, the effects of experience on face recognition are innately constrained (Cohen Kadosh & Johnson 2007). Since our face processing system depends on upright stimuli, such stimuli are 'unmarked', in the sense of cognitively 'natural', and only with natural stimuli can the system readily distinguish 'atypical' (thatcherized) from 'typical' faces. Thus naturalness enhances typicality effects in face recognition.

The closest direct parallel to face recognition in phonology is speech perception, and the conclusions here are similar. For example, click consonants are extremely rare across the world's languages. While they are readily distinguished by adult native speakers of non-click languages, like English, this is because the discrimination of non-native phonemes depends on how they are assimilated by listeners into their native phonemic inventory, and clicks are so different from any English phoneme that English listeners fail to assimilate them at all, forcing them to discriminate

the sounds on a purely acoustic basis (Best & McRoberts 2003). Thus the markedness of clicks affects the degree to which the speech-processing system is engaged, resulting in at least a quantitative, perhaps even qualitative, difference in processing. If English participants were given an acceptability judgment task with nonword items containing clicks, it seems reasonable to speculate that items differing only in click type would be judged as equally un-English-like, even though some clicks (like the lateral click [ǁ]) share more features with English phonemes (like the lateral liquid /l/), and thus are more lexically typical, than others (like the central click [ǀ]).

As far as we know, however, such an experiment has yet to be run. In the next section we test for the expected interaction between lexical typicality and markedness using experimental evidence of a different kind.

3. Markedness and lexical typicality interactions in Mandarin

Our data come from a phonological acceptability judgment experiment run on native speakers of Mandarin, using monosyllabic test items not in the Mandarin syllable inventory but composed of Mandarin phonemic elements. In order to make the results as comprehensive and conclusive as possible, we adopted what is sometimes called a megastudy technique (Balota et al. 2012). In our case, we tested over 3,000 items and over 100 speakers. The experiment produced a large corpus of judgments (http://lngproc.ccu.edu.tw/LDB/) that researchers are free to use to test a wide variety of hypotheses (see e.g. Myers 2015).

There are countless different ways lexical typicality and markedness could be defined, so to keep our discussion focused and theory-neutral, we chose extremely simple definitions, both relating to the onset consonant. Although studies on phonological acceptability often focus on the influence of segment sequences (e.g. Bailey & Hahn 2001), acceptability is also known to be affected by the language-specific frequencies of individual segments (Treiman et al. 2000; Vitevitch et al. 1997), and both language-specific and cross-linguistic segment frequencies affect phonological acquisition (Beckman & Edwards 2010).

Thus in our experiment, lexical typicality was quantified simply in terms of the number of lexical Mandarin syllables that share the test item's onset, while markedness was quantified in terms of the number of languages that have the test item's onset in their phoneme inventory. We describe our methods in more detail in §3.1, and our results in §3.2.

3.1 Methods

Here we review the methods used to generate the corpus of judgments (closely following the description in Myers 2015), as well as the definitions of the independent variables used in the current analysis.

3.1.1 Participants

One hundred and fourteen native speakers of Mandarin Chinese, undergraduates at National Chung Cheng University in southern Taiwan with normal or corrected-to-normal vision, were paid

for their participation. The results for four participants (two males and two females) were excluded because they failed to return for the second session. Thus, the results of 110 participants (56 males and 54 females) were included in the database.

3.1.2 Materials

Stimuli were 3,274 nonlexical syllables, each consisting of a sequence of Zhuyin Fuhao (注音符號) symbols (the dominant phonetic spelling system used in Taiwan) in the prosodically permitted order of (optional) onset consonant, (optional) medial glide, and rime, along with a tone, all written horizontally, as in computer key-in systems. Zhuyin Fuhao distinguishes 22 onsets (including none), four medials (including none), four tones (not counting the mark for tonelessness, not used in this study), and 14 rimes (including none: orthographic convention omits the rime symbol for dento-alveolar sibilants like /s/ and retroflexes like /ʐ/, where the vowel is essentially a sonorant realization of the onset). We did not include the rhotic rime in the experiment, since it never combines with other elements in tautomorphemic syllables, and also excluded symbol strings without segmental content (i.e. just a tone mark). We also excluded forms with non-dento-alveolar sibilant and non-retroflex onsets written without a rime symbol, violating a basic orthographic convention (56 = 4 tones × 14 such onsets; e.g. ㄖˊ /ʐɨ^{35}/ and ㄙˊ /sɨ^{35}/ were tested, but not the unpronounceable ㄅˊ /p^{35}/, ㄉˊ /t^{35}/, ㄍˊ /k^{35}/, or ㄐˊ /tɕ^{35}/). This gave 4,516 (= 22 × 4 × 4 × (14 – 1) – 4 – 56) logically possible Mandarin syllables expressible with Zhuyin Fuhao. Lexical syllables were taken to be the 1,254 listed in Tsai (2000)—excluding 11 toneless and four rhotic rime syllables—along with three recent neologisms (/piaŋ^{51}/, /liaŋ^{55}/, /piaŋ^{55}/). The total number of nonlexical test syllables was thus 3,274 (= 4,516 – 1,254 + 11 + 4 – 3).

Even though our megastudy was designed to study phonological processing, we used written stimuli instead of auditory stimuli for a number of reasons. First, written stimuli can be precisely controlled; auditory stimuli must be produced by some specific speaker or speech synthesis program which may introduce idiosyncrasies. Second, written stimuli can be processed by participants very quickly, shortening trial durations, which allows more items to be tested. Third, and perhaps most importantly, written stimuli encourage a more abstract level for the judgments than merely phonetic processing; this is the logic behind the use of written stimuli in other experimental studies on phonology, like those of Bailey & Hahn (2001), Berent & Lennertz (2010), and Hayes & White (2013).

The lexical typicality of each test item (of those with onsets) was quantified as the number of lexical Mandarin syllables (as defined earlier) that share the test item's onset (so we only counted onset /n/, not coda /n/). These frequencies are thus type frequencies, not token frequencies, reflecting the realization of Mandarin grammar in the lexicon, not the processing of Mandarin phonology in fluent speech. The left side of Table 1 lists the 21 Mandarin onsets in decreasing order of lexical type frequency. In OT terms, one can think of onset frequency as reflecting the Mandarin-specific constraint ranking (perhaps involving many constraints) that favors certain onsets, like /l/, over others, like /ʐ/. We then took the logarithm (here, natural log, based on e = 2.718...) of the frequencies in an attempt to reduce some of the skew that frequency distributions tend to have (Baayen 2008).

Table 1: Lexical and cross-linguistic frequencies for Mandarin onset consonants

Mandarin frequency				UPSID frequency			
l	82	m	56	m	425	tʰ	49
tʰ	67	n	54	k	403	ts	45
x	66	ɕ	52	p	375	x	44
ʈʂʰ	66	kʰ	51	n	202	tsʰ	25
k	60	tɕʰ	49	f	180	ʂ	23
t	59	ts	48	t	152	ʈʂ	16
ʂ	59	tsʰ	46	l	136	tɕ	12
pʰ	57	tɕ	45	s	135	ɕ	11
ʈʂ	57	s	41	kʰ	103	ʈʂʰ	9
p	56	f	31	pʰ	101	ʐ	9
		ʐ	31			tɕʰ	1

The markedness (naturalness) of each test item was quantified using the UCLA Phonological Segment Inventory Database (UPSID; Maddieson 1984), via the free web interface at http://web.phonetik.uni-frankfurt.de/upsid.html. That is, we counted the number of languages in UPSID that share the test item's onset consonant, regardless of syllable position (information not available in UPSID). UPSID is a relatively small database of only 451 languages, but it is designed to reflect linguistic diversity, rather than oversampling from closely related languages. Ideally, then, segments common in UPSID should truly be favored by the human phonological system, not popular merely due to historical accident.

Mandarin happens to be one of the languages in UPSID, so the phonetic descriptions of its phonemes were simply adopted here. For example, Mandarin /t/ is described in UPSID as a 'voiceless dental/alveolar plosive', not merely as a 'voiceless alveolar plosive', which is classed as a different segment with a different language count (181, compared with 152 for Mandarin /t/). The right side of Table 1 lists the 21 Mandarin onsets in decreasing order of cross-linguistic frequency, as previously defined. Again, these frequencies were log-transformed before entering them into the statistical analyses.

Presumably because the forces that underlie cross-linguistic frequencies also influence the development of any specific lexicon, there is a slight positive correlation between log Mandarin onset frequency and log UPSID consonant frequency, though it is far from statistically significant across the 21 onsets ($r = .16$, $t(19) = 0.71$, $p = .49 > .05$). Even if this correlation were statistically reliable, it would only imply that $r^2 = .03$, or 3%, of the variance in one variable is predictable from the other. Thus in the statistical analyses described in the following, we assumed that the two variables are uncorrelated.

3.1.3 Procedure

The experiment was run using E-Prime 2.0 (Schneider et al. 2002). After filling out informed consent forms, which among other things confirmed their right to drop out of the long experiment

without penalty, participants completed some pretests (not relevant here), and then the main experiment began. All participants were visually presented with all 3,274 syllables, written horizontally from left to right in Zhuyin Fuhao, in two blocks (1,599 in the first, 1,675 in the second), each in random order. Each block was preceded by four practice trials to familiarize participants with the task (or refamiliarize them, in the second session). There were rest breaks every 160 trials. All participants were asked to place their right and left hands, respectively, over the right and left sides of the computer keyboard, and then to judge each stimulus as 'like Mandarin' (「像國語」) by pressing the 'L' key (right hand) or as 'not like Mandarin' (「不像國語」) by pressing the 'S' key (left hand), as quickly as possible. Each trial began with the mark '+' for 1,000 milliseconds (one second) to ensure that their eyes were pointed in the right direction when the stimulus appeared, followed by the target syllable for 4,000 ms. Each trial ended when the participant made a response or the 4,000 ms limit passed. The task was split into two blocks because it took participants three hours to complete the pretests and judge all syllables. The second session took place on the same day (after a two-hour lunch break), or on another day (one to 10 days later).

By asking participants to make quick binary judgments of acceptability, the overall experiment was shortened, permitting the testing of more items. Binary judgments don't sacrifice information, since even when given a gradient scale, people tend to give acceptability judgments that cluster at the low and high ends of the scale anyway (Sprouse 2007). Binary judgments still permit the analysis of gradient patterns in the proportions of positive versus negative judgments across participants or items (Cowart 1997) without loss of statistical power (Bader & Häussler 2010; Weskott & Fanselow 2011).

3.2 Results

A total of 1,727 responses were lost for being too slow (over 4,000 ms), and an additional 12,209 were dropped due to their unrealistic reaction times (less than 100 ms, suggesting accidental responses). Though the amount of lost data sounds large, note that there were 360,140 (= 3,274 × 110) possible responses, so the lost data comprised less than 3% of the total. Despite the four-second window for responses, it seems that the instruction to respond quickly, combined with the natural desire to finish the long experiment as soon as possible, led participants to respond about as fast as in a typical word-reading experiment, though with considerable variability (mean 802 ms, standard deviation 587 ms).

In the vast majority of trials, items were rejected as not like Mandarin; only 11% of the valid responses were positive, and one participant even rejected all items. As Myers & Tsay (2005) note, the restrictive phonotactics of Mandarin make new syllable coinages extremely rare, so speakers are reluctant to accept unattested syllables of any sort. Nevertheless, given the large number of items and participants, there were 37,603 positive responses from the 109 remaining participants, with one participant even showing a 65% acceptance rate.

Five test items were rejected by all participants: /xyo^{55}/, /fyou^{55}/, /fyou^{214}/, /ʨɤ^{35}/, and /ʈʂyau^{35}/ (i.e. Zhuyin Fuhao ㄏㄩㄛ, ㄈㄩㄡˇ, ㄈㄩㄡˋ, ㄓㄜˊ, ㄓㄩㄠˊ). The five best-scoring items were /man^{55}/, /uo^{35}/, /ei^{35}/, /ei^{51}/, and /ei^{55}/ (ㄇㄢ, ㄨㄛˊ, ㄟˊ, ㄟˋ, ㄟ), with cross-participant acceptance rates ranging from .45 to .50.

Our focus here on onset consonants led us to drop all onsetless syllables, including glide-initial syllables, leaving 3,187 items. This resulted in the dropping of an additional 11,946 experimental trials (responses to onsetless syllables), representing around 3% of the reduced data set and 8% of the original data set. As it happened, one participant rejected all test items with onsets, reducing our informative participants to 108.

Because our primary goal was to test for not merely main effects of lexical typicality and markedness, but also their interaction, we standardized both of these variables, by subtracting their means (which makes the new means zero) and dividing by their standard deviations (which gives their distributions the same magnitude). The *z* scores produced by this transformation ensure that the main effects will tend to remain as they would even if no interaction had been tested; it also puts all regression coefficients on the same standardized scale (Aiken & West 1991).

The statistical analysis used a powerful technique called mixed-effects logistic regression (Baayen 2008). Mixed-effects regression is so called because it can simultaneously tease apart the influence of random effects (e.g. participants and test items) and fixed effects (in this case, the two types of onset frequencies). Analyses were run using the lme4 package (Bates et al. 2014) in the R statistical programming language (R Core Team 2014). Since participants may respond systematically differently to different types of items, Barr et al. (2013) recommend that this random-effect interaction also be taken into consideration, by nesting the fixed factors and their interaction within the participants. This type of analysis is computationally intensive and prone to crashing, but fortunately in this case it was able to build the statistical model successfully (although it took over four hours to do so).

The statistical results are shown in Table 2. The large negative standardized coefficient (β) for the intercept merely indicates that participants had an overall tendency to reject items. All of the other coefficients are positive, suggesting facilitation of acceptability. Thus there are positive effects of both lexical typicality (log onset Mandarin type frequency *z* scores) and markedness in its positive sense of naturalness (log consonant frequency in UPSID *z* scores). Crucially, there is also a positive (enhancing) interaction between the two factors. All of these effects are highly significant.

Table 2: Results of mixed-effects logistic regression

	β	*SE*	*z*	*p*	
Intercept	–3.558	0.188	–18.933	< .001	*
Mand freq	0.168	0.024	6.935	< .001	*
UPSID freq	0.115	0.027	4.192	< .001	*
Mand freq × UPSID freq	0.123	0.028	4.432	< .001	*

Notes: * represents statistical significance ($p < .05$). β, *SE*, *z*, *p* represent, respectively, the standardized regression coefficient, the standard error, the *z* score for the coefficient (β/SE), and the *p* value (computed using *z*). Mand freq and UPSID freq represent, respectively, log onset Mandarin type frequency *z* scores and log consonant UPSID frequency *z* scores.

To get a more concrete sense of the effect of the two variables and their interaction, we turn to Figure 1. In each scatterplot, the dots represent the test items, and the y-axis represents

cross-participant acceptance transformed into log odds (the natural log of the ratio of acceptance to rejection), which is the scale used by logistic regression (hence the name). We had to remove those five items rejected by all participants, since the log of zero is undefined. The values are negative because items tended to be rejected more than accepted, but higher scores still represent greater acceptability. The x-axes represent the three key model parameters listed in Table 2: lexical typicality, markedness, and their interaction. As in the regression model itself, the interaction is quantified as the product of the other two parameters, so the right end of the x-axis shows items that are both lexically typical and unmarked (product of two positive *z* scores) or both atypical and marked (product of two negative *z* scores), while the left end shows items that are typical but marked or atypical but unmarked (product of *z* scores of opposite sign). The trend lines (locally weighted scatterplot smoothing lines) reflect local fluctuations in the data.

As implied by Table 2, both lexical typicality and markedness (in its positive sense of naturalness) have positive effects on acceptability (overall rising slopes), and crucially, so does their interaction: acceptability tends to be higher for items that have the same sign for typicality and markedness (both positive or both negative). The overall rising slope of the interaction trend line can be understood as an increase in the lexical typicality effect as a function of increasing naturalness (or as an increase in the markedness effect as a function of increasing lexical typicality).

Figure 1 also demonstrates that the distributions of the log odds scores are roughly symmetrical, with all trend lines far from the 'floor' (important because illusory interactions can arise through a so-called floor effect; Kang & Waller 2005).

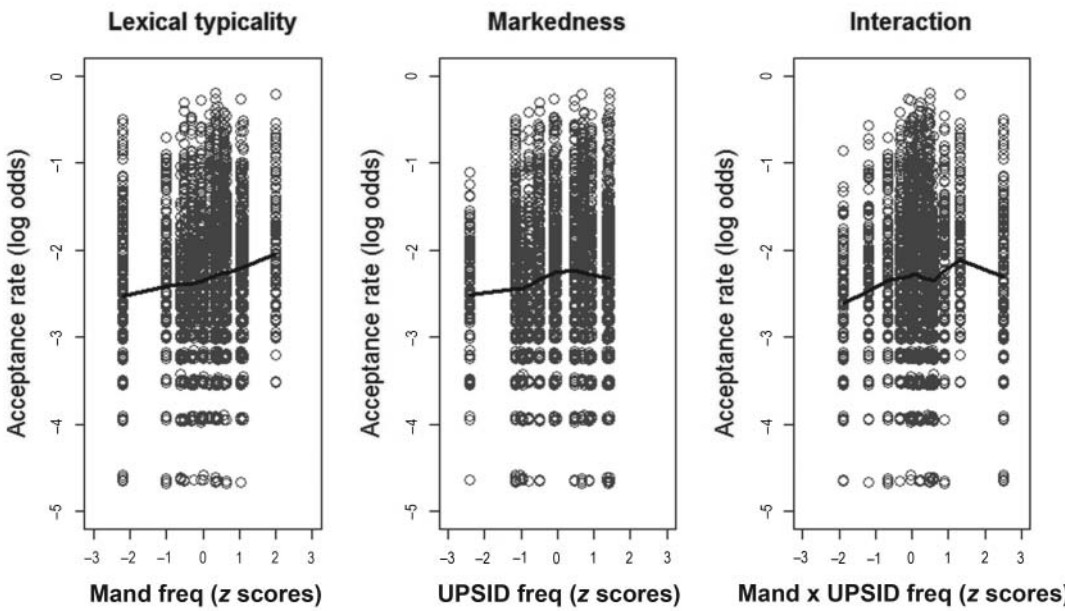

Figure 1: Markedness and lexical typicality effects on Mandarin acceptability judgments

While all three trend lines have overall positive slopes, only lexical typicality has a truly linear relationship with acceptability. The weakened influence of higher levels of naturalness on

acceptability (i.e. the flattened trend line on the right end of the center scatterplot) may suggest that speakers are more sensitive to the violation of constraints than to obedience to them, consistent with the intrinsically negative nature of OT markedness constraints. The sudden drop on the right end of the interaction trend line seems to be accidental, however, since it is caused by the single consonant /ʐ/ (the rightmost column of dots).

Such nonlinearities are intriguing but should not distract from our key finding: lexical typicality and markedness interact. Moreover, as expected from general cognitive considerations but not from standard OT, this interaction involves the enhancement of the lexical typicality effect for less marked (more natural) forms.

4. Markedness and lexical typicality interactions in English

Since the claim addressed in this article should apply universally, we decided to look at data in a second language. As we saw earlier, Hayes & White (2013) report a UG experiment in which English participants judged the acceptability of nonwords designed to obey or violate either natural or unnatural phonotactic constraints generated by their HG learning algorithm. Although they never test the interaction between lexical typicality and markedness, their published results can readily be reanalyzed to do so.

Unlike our database-derived quantification of markedness, Hayes & White (2013) determine the naturalness of their constraints solely by expert opinion, but as we saw with their anti-coda-glide constraint (natural) and anti-word-initial-/u,ʊ/ constraint (unnatural), their opinions seem sound. Meanwhile, lexical typicality is precisely what their HG learning algorithm is designed to compute, although the authors don't express it this way. Specifically, the algorithm generates a set of weighted phonotactic constraints, where each weight represents the degree to which the constraint is obeyed in the English lexicon. Thus an item that violates a constraint with a higher weight is less lexically typical than an item that violates a lower-weight constraint. In an appendix (pp. 70–71), the authors give the weight and naturalness status of the constraint associated with each test item, along with its mean judgment score (on a log-transformed continuous scale).

For each of the 80 test items (10 natural and 10 unnatural constraints, with two test pairs per constraint), we coded it for obedience to the relevant constraint, the naturalness of this constraint (1 = natural, –1 = unnatural), the HG weight of this constraint (as *z* scores), and the mean rating. Since items came in pairs, we converted their ratings into one difference score per pair by subtracting the rating for the disobedient item from that for its matched obedient control (which violates no constraint not also violated by the disobedient item). The worse the disobedient item relative to its matched control, the higher the difference score, indicating a greater effect of the tested constraint on acceptability. Thus if lexical typicality has a positive effect on acceptability, higher constraint weights (indicating stronger lexical typicality) should be associated with higher difference scores.

The resulting scatterplot for the 40 test item pairs is shown in Figure 2, with separate linear trend lines for the pairs associated with natural versus unnatural constraints. Although there is considerable noise in the data (indicated by how far the data points are from the trend lines), the plot demonstrates three key patterns. First, consistent with the conclusions of Hayes & White (2013), scores for natural constraints are higher than unnatural constraints (i.e. the solid line is above the

dashed line). Second, the overall effect of lexical typicality (constraint weight) on acceptability also seems to be positive (i.e. an imaginary line equidistant between the solid and dashed lines, representing their average, would have a rising slope). Third, and most important for our purposes, the positive effect of lexical typicality on acceptability is only seen for item pairs testing natural constraints (i.e. only the solid line has a rising slope). This third observation represents the interaction between markedness and lexical typicality. The plot thus suggests that the results of this English study are very similar to what we observed in our Mandarin study: both lexical typicality and markedness improve acceptability, and lexical typicality effects are stronger in less marked contexts.

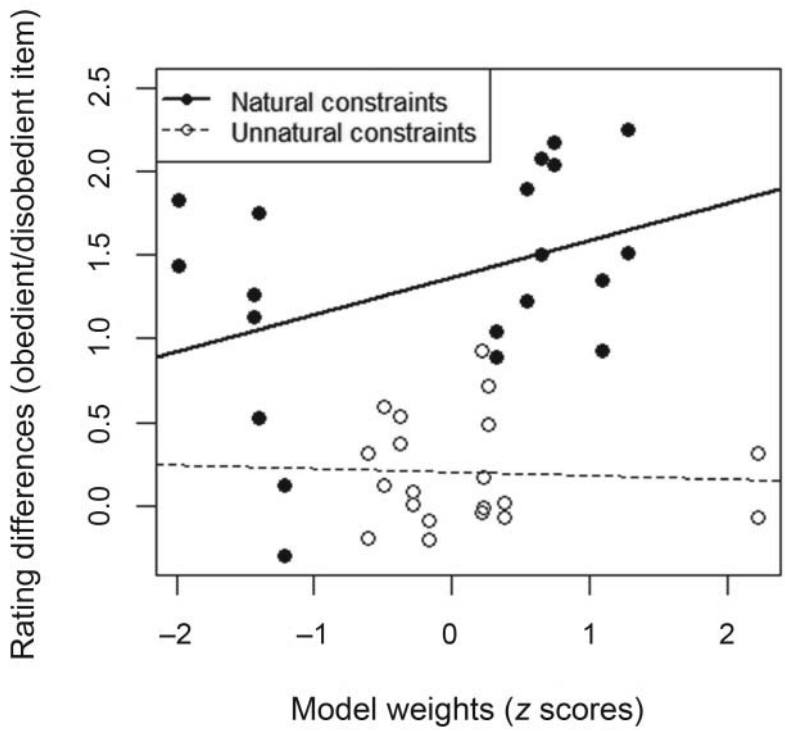

Figure 2: Reanalysis of Hayes & White (2013)

Unfortunately, given that the materials were designed only to test markedness, not lexical typicality or its interaction with markedness, and given that we only have access to a small sample of summary data, the impressions given by the plot cannot be backed up statistically. Using a multiple regression model with item pair as the random variable, only naturalness has a significant effect ($\beta = 0.58$, $SE = 0.08$, $t(36) = 7.12$, $p < .001$); neither the main effect of lexical typicality (constraint weight: $\beta = 0.10$, $SE = 0.09$, $t(36) = 1.14$, $p = .26$) nor the interaction ($\beta = 0.12$, $SE = 0.09$, $t(36) = 1.36$, $p = .18$) reach statistical significance.

Ironically, however, precisely because the authors themselves made no explicit attempt to test for interactions, the fact that there nevertheless seems to be a tendency for just the expected interaction remains highly suggestive, and certainly worthy of follow-up in new experiments on English and other languages.

5. Formalization in Optimality Theory

If standard OT cannot handle the kind of interaction that we expect, and in fact observe, how might it be revised to do so? In this section we suggest one possibility: the interaction between lexical typicality and markedness may arise from interactions among the markedness constraints themselves. We first justify taking a formal approach, rather than adopting a purely psycholinguistic explanation, in §5.1. We explain the core formal intuition in §5.2, showing, however, that the most widely used form of constraint interaction—local conjunction—does not give the desired results. This motivates the adoption of a different kind of constraint interaction into the repertoire of OT devices, as we argue in §5.3.

5.1 Why formalize?

If the interaction between lexical typicality and markedness can only be detected in psycholinguistic experiments, and if it is truly akin to extra-linguistic phenomena like the Thatcher face illusion, then why bother with a formal OT analysis at all? The reason is simple. Not only does being formal merely mean being precise, something that all sciences strive for, but any potential extra-grammatical explanation must be made formally precise as well, and it is not yet clear how this could be done.

An interesting near-miss comes from psychophysics, which searches for formal laws relating physical stimuli to percepts. One of the most famous of these is the power law of Stevens (1957), which implies, among other things, that the perceived change in a physical magnitude depends on the overall magnitude, not just on the change size. This is why it is easy to notice the lighting of a single candle in a dark room, but not in a bright room.

This law also seems to apply to acceptability judgments, at least in syntax, where it was tested by Keller (2003). He found that the difference in acceptability between German sentences with one versus two violations was much larger than the difference for sentences with four versus five violations. Since the total number of violations represents markedness, the power law implies that a small difference in markedness will be more readily perceived if the test items are themselves less marked.

Unfortunately, Keller (2003) is merely describing a nonlinearity in raw acceptability, not an interaction per se. Indeed, the only reason why the trend line in the leftmost scatterplot in Figure 1 appears linear is that we transformed acceptability on a log scale, and the log and power functions are inverses of each other; the raw acceptability rates actually differ slightly more among the more lexically typical items than among the less lexically typical items, consistent with Keller's application of the power law. (The nonlinearity in the markedness trend line in the center scatterplot of Figure 1 does not conform to the power law, however.)

Nevertheless, the key observation in our study does not relate to (non)linearities in any individual variable, but rather to the interaction of lexical typicality with markedness that is reflected in the rightmost scatterplot in Figure 1. This interaction was untested in Keller (2003): his sentences differed not only in overall markedness, but also in typicality (relative to the participants' mental corpora of German sentences), so the two factors were fully conflated.

Thus while it is undeniable that extra-grammatical models of our interaction are worth looking for, an attempt to formalize it now may actually aid in this search. This is what we do below.

5.2 Constraint conjunction

In §2.3 we saw that an OT grammar is mathematically equivalent to a linear equation. Because the effects of the markedness constraints are simply added together, the effect of one doesn't modulate the effect of the others. This is why standard OT predicts that markedness and lexical typicality should not interact, even though it seems that they do in reality.

In statistics, interactions are represented as the (multiplicative) product of the two interacting factors. For example, if constraints A and B interact, then the weight Weight_{AB} of their product, as in the equation in (9), should be significantly different from zero.

(9) $\text{Harmony} = \text{Weight}_A \times A + \text{Weight}_B \times B + \text{Weight}_{AB} \times AB$

Interestingly, given its historical and mathematical links with regression, standard OT actually has a device for representing constraint interactions in a similar way. In local constraint conjunction (Smolensky 1993, 2006), a coordinated constraint is violated if and only if the two constraints composing it share the same phonological domain (e.g. same segment or same syllable) and are both violated. If the locally conjoined constraint is ranked at least as high as its component constraints, it 'bans only the worst of the worst' (for which Smolensky 2006:43 coins the amusing acronym BOWOW).

Local constraint conjunction has been widely used in the OT literature (see references in Łubowicz 2005), and indeed we have already seen it used by Coetzee (2008). In his analysis, the odd looking *spVp constraint in English is a locally conjoined constraint derived from two simpler markedness constraints, namely *sp and *pVp, conjoined within the syllable (σ). As sketched below in (10), English permits any combination of *sp and *pVp violations except for their simultaneous violation: **spop* is banned, but *span* and *pipe* are both permitted. The requirement that the two component constraints share a domain is crucial (Łubowicz 2005): *sp $\&_\sigma$ *pVp should not ban the compound *pipe span*.

Unfortunately, Smolensky's local constraint conjunction results in the wrong type of interaction. This can be seen in (11) in a schematic example with the abstract markedness constraints A, B, C. Here, B $\&_D$ C is a locally conjoined constraint in some domain D that is violated if and only if both B and C are violated.

(10)

Input	Output	*sp $\&_\sigma$ *pVp	FAITH	*sp	*pVp
/spɑp/	[spɑp]	*		*	*
	☞ [stɑp]		*		
/spæn/	☞ [spæn]			*	
	[stæn]		*		
/paɪp/	☞ [paɪp]				*
	[taɪp]		*		

(11) a. Less marked items

Weights:		8	4	2	1
		B &$_D$ C	A	B	C
–4	aBC		*		
–2	AbC			*	
–1	ABc				*

b. More marked items

Weights:		8	4	2	1
		B &$_D$ C	A	B	C
–6	abC		*	*	
–5	aBc		*		*
–11	Abc	*		*	*

The variability in the harmony scores for the less marked items in (11a) is less than the variability for the marked items in (11b), both in range (3 = –1 – (–4) versus 6 = –5 – (–11)) and in standard deviation (*SD* = 1.53 versus *SD* = 3.21). In other words, if we use local constraint conjunction to capture the interaction between constraints, we actually predict exactly the opposite of what we observe in both the Mandarin and English acceptability judgments.

Potts et al. (2010) argue that BOWOW phenomena can be modeled more effectively by eliminating the restrictions on OT constraint weights, that is, by generalizing OT to HG. This is because HG is capable of weighting lower-weight constraints so that they can 'gang up' on higher-weight constraints. For example, if we give both *sp and *pVp the weight of 2 and FAITH the weight of 3, each individual markedness constraint is too weak to override FAITH (2 < 3) but together they can (2 + 2 > 3).

Regardless how well HG models BOWOW cases in general, it fails to capture our observed markedness and lexical typicality interactions. This is because HG has no interaction factors at all, so just like standard OT, it predicts that the two factors should be independent. This is illustrated schematically in (12), which shows tables identical to those given earlier in (7), except with weights designed to permit the ganging up of constraints B and C over constraint A. The variability in harmony scores for both less marked and more marked items is exactly the same, both in range (1 = –2 – (–3) versus 1 = –4 – (–5)) and in standard deviation (*SD* = 0.58 versus *SD* = 0.58).

(12) a. Less marked items

Weights:		3	2	2
		A	B	C
–3	aBC	*		
–2	AbC		*	
–2	ABc			*

b. More marked items

Weights:		3	2	2
		A	B	C
–5	abC	*	*	
–5	aBc	*		*
–4	Abc		*	*

5.3 Conjunctive coordination

Although neither local constraint conjunction nor gang effects in HG capture the observed enhancement of lexical typicality effects by markedness, this doesn't rule out formalization via some other type of constraint interaction. As Crowhurst & Hewitt (1997) point out, we can define a larger family of constraint interactions by using basic logical operations. If obeying a constraint is interpreted as 'true' and violating a constraint as 'false', the familiar local constraint conjunction of Smolensky (1993, 2006) is actually equivalent to disjunction, not conjunction, since the constraint A & B is true (obeyed) if and only if A *or* B is obeyed (i.e. A $\vee$ B in logical notation). The typology presented by Crowhurst & Hewitt (1997) then adds constraints derived via the logical operations of conjunction (*and*, symbolized $\wedge$) and implication (*if-then*, symbolized $\Rightarrow$).

They focus particularly on what they call conjunctive coordination ($\wedge$), in which the coordinated constraint is *obeyed* if and only both of its component constraints are *obeyed*. The behaviors of disjunctively and conjunctively coordinated constraints are contrasted in (13), where A and B represent constraints, an empty cell is interpreted as 'true', and a starred cell is interpreted as 'false'. As usual, coordinated constraints must share a domain D.

(13) a. Disjunctive coordination (local constraint conjunction; Smolensky 1993)

A	B	A $\vee_D$ B
	*	
*		
*	*	*

b. Conjunctive coordination (Crowhurst & Hewitt 1997)

A	B	A $\wedge_D$ B
	*	*
*		*
*	*	*

Conjunctive coordination turns out to capture just the kind of interaction between lexical typicality and markedness that we observe, as shown in the schematic examples in (14). As desired,

conjunctively coordinating the markedness constraints B and C yields harmony scores for less marked items that show more variability than for more marked items, both in range (6 = –4 – (–10) versus 3 = 11 – (–14)) and in standard deviation (*SD* = 3.21 versus *SD* = 1.53). The algebraic reason for this success is that a conjunctively coordinated constraint is violated even if only one of its component constraints is violated, thus reducing the difference in harmony across forms that violate one versus both component constraints. This reduction is greater the more coordinated constraints are violated, which is a function of a form's overall markedness.

(14) a. Less marked items

Weights:		8	4	2	1
		B $\wedge_D$ C	A	B	C
–4	aBC		*		
–10	AbC	*		*	
–9	ABc	*			*

b. More marked items

Weights:		8	4	2	1
		B $\wedge_D$ C	A	B	C
–14	abC	*	*	*	
–13	aBc	*	*		*
–11	Abc	*		*	*

Given the neglect of conjunctive coordination in the OT literature, however, we must ask whether adding it to the inventory of OT devices has any unintended bad consequences. Wolf (2007) even takes the trouble to argue against all possible types of constraint coordination other than Smolensky's familiar version. He starts by pointing out that tables like those in (13) actually predict 16 different types of constraint coordination, corresponding to the 2^4 different patterns of starred and empty cells in the column under the coordinated constraint. He then goes on to show that if any OT constraint, whether simple or derived, must be either a markedness or faithfulness constraint, then only local constraint conjunction (disjunction) is guaranteed to conform to this principle regardless of what types of constraints are coordinated. However, since our focus here is on the coordination of markedness constraints, which necessarily yields a markedness constraint, Wolf's critique may not be relevant here.

There is also the question of how conjunctively coordinated constraints could be learned. Smolensky (2006) admits that even for his local constraint conjunction (i.e. disjunctive coordination), learning algorithms have yet to be developed and tested. Related algorithms have been proposed, however. Hayes & Wilson (2008) and Pater (2014) describe HG learning models that use lexical experience to induce not just constraint weights, but also some of the constraints themselves, banning particular feature combinations (as in the English constraints of Hayes & White 2013 discussed earlier in §2.2). In essence, their algorithms posit feature co-occurrence constraints for all

features that appear in lexical forms, and then use ordinary HG learning algorithms to set the weights, so that useless constraints are demoted to oblivion. Clearly, feature co-occurrence constraints like *[+F, +G]$_D$ are equivalent to locally conjoined constraints like *[+F] &$_D$ *[+G], being violated if and only if both [+F] and [+G] are present. Unfortunately, Hayes & Wilson (2008) are forced to use various tricks to search efficiently through all logically possible feature combinations, which they estimate at around 3.5 million for the English lexicon (p. 392, fn. 8); Pater (2014) admits that he avoids this problem only by artificially limiting his feature set. Learning conjunctively coordinated constraints faces the same computational challenge.

Whether such formal issues have real consequences remains to be seen, but at least we have shown that conjunctively coordinated constraints work better than both standard OT and OT with disjunctively coordinated constraints at describing the observed enhancing interaction between lexical typicality and markedness.

6. Concluding remarks

Though neither our empirical data nor our analysis of them dispel all doubts, we feel we have made a good case that markedness (naturalness) enhances the effect of lexical typicality on phonological acceptability judgments, that this phenomenon should be observed wherever it is looked for, that its source is within the grammar and not the processing, that it derives specifically from the interaction among individual markedness constraints, and that OT, supplemented with conjunctive coordination of constraints, can formalize it.

However, the reader may have noticed a large gap between the types of empirical data we discuss and our analytical framework. In particular, even though our formal analysis depends on markedness constraints that refer to specific phonological properties, we have said nothing about what these properties actually are for our Mandarin data (compare Hayes & White 2013, who explicitly state the constraints tested in their English study). Our Mandarin constraints presumably relate to features of the onsets, but the acceptability of a syllable obviously depends on much more than just the onset. A given onset may be favored or disfavored in our judgment data not solely because of the consonant itself, but because of the phonotactic contexts in which it appears. Addressing such issues will require careful empirical and formal work. For example, even if it is discovered that two markedness constraints fail to interact, our proposal isn't challenged unless they share a phonological domain in the relevant test items, as coordinated constraints must do.

More generally, further analysis is needed, involving not just data from different languages, but also formalization using different and perhaps more sophisticated quantifications of markedness and lexical typicality than those we use here. Even if our empirical claims prove robust, alternative analyses should be considered beyond the one we have proposed, relying as it does on the relatively untested device of conjunctive coordination. Nevertheless, we hope that we have convinced the reader that it is fruitful for phonologists, both theoretical and experimental, to consider markedness and lexical typicality together, not just as separate phenomena.

References

Aiken, Leona S., & Stephen G. West. 1991. *Multiple Regression: Testing and Interpreting Interactions*. London: SAGE.

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26.1:9–41.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge & New York: Cambridge University Press.

Bader, Markus, & Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46.2:273–330.

Bailey, Todd. M., & Ulrike Hahn. 2001. Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44.4:568–591.

Balota, David A., Melvin J. Yap, Keith A. Hutchison, & Michael J. Cortese. 2012. Megastudies: what do millions (or so) of trials tell me about lexical processing? *Visual Word Recognition*, Vol. 1: *Models and Methods*, *Orthography and Phonology*, ed. by James S. Adelman, 90–115. Hove: Psychology Press.

Bard, Ellen G., Dan Robertson, & Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72.1:32–68.

Barr, Dale J., Roger Levy, Christoph Scheepers, & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* 68.3:255–278.

Bartlett, James C., & Jean Searcy. 1993. Inversion and configuration of faces. *Cognitive Psychology* 25.3:281–316.

Bates, Douglas, Martin Maechler, Ben Bolker, & Steven Walker. 2014. lme4: Linear mixed-effects models using Eigen and S4. http://arxiv.org/abs/1406.5823.

Becker, Michael, & Kathryn F. Potts. 2011. The emergence of the unmarked. *The Blackwell Companion to Phonology*, Vol. 3: *Phonological Processes*, ed. by Marc van Oostendorp et al., 1363–1379. Oxford: Wiley-Blackwell.

Beckman, Mary E., & Jan Edwards. 2010. Generalizing over lexicons to predict consonant mastery. *Laboratory Phonology* 1.2:319–343.

Berent, Iris, & Tracy Lennertz. 2010. Universal constraints on the sound structure of language: phonological or acoustic? *Journal of Experimental Psychology: Human Perception and Performance* 36.1:212–223.

Berent, Iris, & Joseph Shimron. 1997. The representation of Hebrew words: evidence from the obligatory contour principle. *Cognition* 64.1:39–72.

Berent, Iris, Donca Steriade, Tracy Lennertz, & Vered Vaknin. 2007. What we know about what we have never heard: evidence from perceptual illusions. *Cognition* 104.3:591–630.

Berent, Iris, Colin Wilson, Gary F. Marcus, & Douglas K. Bemis. 2012. On the role of variables in phonology: remarks on Hayes and Wilson 2008. *Linguistic Inquiry* 43.1:97–119.

Berger, Adam L., Stephen A. Della Pietra, & Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22.1:39–71.

Best, Catherine C., & Gerald W. McRoberts. 2003. Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech* 46.2–3:183–216.

Boersma, Paul, & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32.1:45–86.

Chomsky, Noam, & Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1.2:97–138.

Coetzee, Andries W. 2008. Grammaticality and ungrammaticality in phonology. *Language* 84.2: 218–257.

Cohen Kadosh, Kathrin, & Mark H. Johnson. 2007. Developing a cortex specialized for face perception. *TRENDS in Cognitive Sciences* 11.9:367–369.

Coleman, John, & Janet B. Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*, 49–56. Somerset: Association for Computational Linguistics.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. London: SAGE.

Crowhurst, Megan, & Mark Hewitt. 1997. Boolean operations and constraint interactions in Optimality Theory. Manuscript. University of North Carolina at Chapel Hill & Brandeis University.

Davis, Stuart M. 1989. Cross-vowel phonotactic constraints. *Computational Linguistics* 15.2:109–110.

Frisch, Stefan A., & Bushra A. Zawaydeh. 2001. The psychological reality of OCP-Place in Arabic. *Language* 77.1:91–106.

Golston, Chris. 1996. Direct Optimality Theory: representation as pure markedness. *Language* 72.4:713–748.

Greenberg, Joseph H., & James J. Jenkins. 1964. Studies in the psychological correlates of the sound system of American English. *Word* 20.2:157–177.

Hahn, Ulrike, & Todd M. Bailey. 2005. What makes words sound similar? *Cognition* 97.3:227–267.

Hayes, Bruce, Péter Siptár, Kie Zuraw, & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85.4:822–863.

Hayes, Bruce, & James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44.1:45–75.

Hayes, Bruce, & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.3:379–440.

Kang, Sun-Mee, & Niels G. Waller. 2005. Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement* 29.2:87–105.

Kawahara, Shigeto. 2011. Experimental approaches in theoretical phonology. *The Blackwell Companion to Phonology*, Vol. 4: *Phonological Interfaces*, ed. by Marc van Oostendorp et al., 2283–2303. Oxford: Wiley-Blackwell.

Kawahara, Shigeto, & Sophia Kao. 2012. The productivity of a root-initial accenting suffix, [-zu]: judgement studies. *Natural Language & Linguistic Theory* 30.3:837–857.

Keller, Frank. 2003. A psychophysical law for linguistic judgments. *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, ed. by Richard Alterman & David Kirsh, 652–657. Boston: Cognitive Science Society.

Keller, Frank. 2006. Linear Optimality Theory as a model of gradience in grammar. *Gradience in Grammar: Generative Perspectives*, ed. by Gisbert Fanselow, Caroline Féry, Ralf Vogel & Matthias Schlesewsky, 270–287. Oxford & New York: Oxford University Press.

Kirby, James P., & Alan C. L. Yu. 2007. Lexical and phonotactic effects on wordlikeness judgments in Cantonese. *Proceedings of the 16th International Congress of Phonetic Sciences*, ed. by Jürgen Trouvain & William J. Barry, 1389–1392. Dudweiler: Pirrot.

Kirchner, Robert. 1997. Contrastiveness and faithfulness. *Phonology* 14.1:83–111.
Leben, William R. 1973. *Suprasegmental Phonology*. Cambridge: MIT dissertation.
Legendre, Géraldine, Yoshiro Miyata, & Paul Smolensky. 1990. Harmonic Grammar: a formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations. *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, ed. by Morton Ann Gernsbacher & Sharon J. Derry, 388–395. Mahwah: Lawrence Erlbaum Associates.
Łubowicz, Anna. 2005. Locality of conjunction. *Proceedings of the 24th West Coast Conference on Formal Linguistics*, ed. by John Alderete, Chung-hye Han & Alexei Kochetov, 254–262. Somerville: Cascadilla Proceedings Project.
Maddieson, Ian. 1984. *Patterns of Sounds*. Cambridge & New York: Cambridge University Press.
Malouf, Robert. 2002. A comparison of algorithms for maximum entropy parameter estimation. *Proceedings of the Sixth Conference on Computational Natural Language Learning* (CoNLL-2002), 49–55. New Brunswick: Association for Computational Linguistics.
McCarthy, John J. 1979. *Formal Problems in Semitic Phonology and Morphology*. Cambridge: MIT dissertation.
McCarthy, John J., & Alan Prince. 1994. The emergence of the unmarked: optimality in prosodic morphology. *Proceedings of the 24th Annual Meeting of the North East Linguistic Society* (*NELS 24*), 333–379. Amherst: GLSA.
Mendoza-Denton, Norma, Jennifer Hay, & Stefanie Jannedy. 2003. Probabilistic sociolinguistics: beyond variable rules. *Probabilistic Linguistics*, ed. by Rens Bod, Jennifer Hay & Stefanie Jannedy, 97–138. Cambridge: MIT Press.
Moreton, Elliott, & Joe Pater. 2012. Structure and substance in artificial-phonology learning, Part II: Substance. *Language and Linguistics Compass* 6.11:702–718.
Myers, James. 2015. Stuck in the middle: Mandarin medials in articulation, parsing, and association. *Capturing Phonological Shades Within and Across Languages*, ed. by Yuchau E. Hsiao & Lian-Hee Wee, 101–119. Cambridge: Cambridge Scholars Publishing.
Myers, James, & Jane Tsay. 2005. The processing of phonological acceptability judgments. *Proceedings of Symposium on 90–92 NSC Projects*, 26–45. Taipei: National Science Council.
Ohala, John J., & Manjari Ohala. 1986. Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. *Experimental Phonology*, ed. by John J. Ohala & Jeri J. Jaeger, 239–252. Orlando: Academic Press.
Pater, Joe. 2014. Canadian raising with language-specific weighted constraints. *Language* 90.1: 230–240.
Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt, & Michael Becker. 2010. Harmonic Grammar with linear programming: from linear systems to linguistic typology. *Phonology* 27.1: 77–117.
Prince, Alan, & Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden: Blackwell.
R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Schneider, Walter, Amy Eschmann, & Anthony Zuccolotto. 2002. *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.
Selkirk, Elizabeth O. 1982. The syllable. *The Structure of Phonological Representations*, Vol. 2, ed. by Harry van der Hulst & Norval Smith, 337–383. Dordrecht: Foris.

Smolensky, Paul. 1993. Harmony, markedness, and phonological activity. Talk presented at Rutgers Optimality Workshop I, New Brunswick, NJ, USA.
Smolensky, Paul. 2006. Optimality in phonology II: harmonic completeness, local constraint conjunction, and feature domain markedness. *The Harmonic Mind: From Neural Computation to Optimality-theoretic Grammar*, Vol. 2: *Linguistic and Philosophical Implications*, ed. by Paul Smolensky & Géraldine Legendre, 27–160. Cambridge: MIT Press.
Sprouse, Jon. 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics* 1:123–134.
Stevens, Stanley S. 1957. On the psychophysical law. *Psychological Review* 64.3:153–181.
Stockall, Linnaea, Andrew Stringfellow, & Alec Marantz. 2004. The precise time course of lexical activation: MEG measurements of the effects of frequency, probability, and density in lexical decision. *Brain and Language* 90.1–3:88–94.
Tesar, Bruce, & Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29.2:229–268.
Thompson, Peter. 1980. Margaret Thatcher: a new illusion. *Perception* 9.4:483–484.
Treiman, Rebecca, Brett Kessler, Stephanie Knewasser, Ruth Tincoff, & Margo Bowman. 2000. English speakers' sensitivity to phonotactic patterns. *Acquisition and the Lexicon*, ed. by Michael B. Broe & Janet B. Pierrehumbert, 269–282. Cambridge & New York: Cambridge University Press.
Tsai, Chih-Hao. 2000. Mandarin syllable frequency counts for Chinese characters. http://technology.chtsai.org/syllable/
Vitevitch, Michael S., & Paul A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40.3:374–408.
Vitevitch, Michael S., Paul A. Luce, Jan Charles-Luce, & David Kemmerer. 1997. Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech* 40.1:47–62.
Wang, H. Samuel. 1998. An experimental study on the phonotactic constraints of Mandarin Chinese. *Studia Linguistica Serica*, ed. by Benjamin K. T'sou, 259–268. Hong Kong: Language Information Sciences Research Center, City University of Hong Kong.
Weskott, Thomas, & Gisbert Fanselow. 2011. On the informativity of different measures of linguistic acceptability. *Language* 87.2:249–273.
Wolf, Matthew. 2007. What constraint connectives should be permitted in OT? *Papers in Theoretical and Computational Phonology*, ed. by Michael Becker, 151–179. Amherst: GLSA.
Zimmer, Karl E. 1969. Psychological correlates of some Turkish morpheme structure conditions. *Language* 45.2:309–321.
Zuraw, Kie. 2007. The role of phonetic knowledge in phonological patterning: corpus and survey evidence from Tagalog infixation. *Language* 83.2:277–316.

[Received 18 September 2014; revised 27 December 2014; accepted 27 December 2014]

Graduate Institute of Linguistics
National Chung Cheng University
Min-Hsiung, Chiayi 621, Taiwan
Lngmyers@ccu.edu.tw

標記性與詞彙典型性：漢語接受度判斷實驗

麥　傑
國立中正大學

眾所皆知，越是接近典型詞的非真詞越容易被說母語者判斷爲可接受；而從結構的標記性來看，則是越少標記的結構越容易被判斷爲可接受。本研究用實驗的方法蒐集了漢語母語人士對非詞彙音節的判斷，建構了巨量資料庫，並首次以此探討標記性與詞彙典型性的交互作用。我們以漢語詞彙音節所共享該實驗項目聲母的程度來定義詞彙典型性；標記性則以該子音分布於跨語言音素系統的多寡來定義。與先前的研究一致，標記性和詞彙典型性都能增進非真詞的接受度，但兩者也確實呈現交互作用：詞彙典型性對於低標記性結構的影響高於對高標記性結構的影響。相同的交互作用也見於對英語的研究。然而這樣的交互作用優選論無法預測，但我們可以用並列結構來解釋。如果我們反向看大家所熟悉的聯合制約，我們會斷定當兩個子限制條件都滿足時，其並列的聯合限制條件才能得到滿足。

關鍵詞：標記性，接受度判斷實驗，漢語，英語，優選論