Phonotactic probabilities in Italian simplex and complex words: a fragment priming study

Giulia Bracco
Università di Salerno
Via Giovanni Paolo II 132
Fisciano (SA)
gcbracco@unisa.it

Basilio Calderone
CNRS & Université de Toulouse II
5 allées Antonio Machado
Toulouse
basilio.calderone@univ.tlse2.fr

Chiara Celata
Scuola Normale Superiore
P.zza dei Cavalieri 7
Pisa
celata@sns.it

1 Introduction

Phonotactics refers to the sequential organization of phonological units that are legal in a language (Crystal 1992). However, legal sound sequences do not all occur with the same probability in a language. *Phonotactic probability* is most often measured in terms of transitional probabilities (TPs) of biphones and has been shown to influence a large range of processes, including infants' discrimination of native language sounds, adults' ratings of the wordlikeness of nonwords (Vitevitch et al. 1997), speech segmentation (Pitt & McQueen 1998, Mattys & Jusczyk 2001), word acquisition (Storkel 2001) and recognition (Luce & Large 2001). Specifically, in the domain of word recognition, high TPs facilitate word and nonword identification in speeded same-different matching tasks, but slow down identification in lexical decision tasks due to the inhibitory effects of a large neighborhood (e.g. Vitevitch & Luce 1999, Luce & Large 2001). Most of the studies on the role of TPs in speech production and perception have been conducted on English.

In this paper we focus on the role of phonotactic probabilities in priming morphologically simplex and complex words in Italian. We investigate whether biphone TPs affect the recognition of word targets after exposure to fragment primes differing in the probability with which the fragment-final consonant predicts the consecutive segment in the target.

We opted for a non-factorial, regression design including lexical and sub-lexical frequency and distributional variables as predictors (see Baayen 2010). In this paper, we report on the results of the study on simplex words only; we however discuss the implications of the current findings for the processing of complex words.

2 Experiment

2.1 Materials and procedure

Forty-two native Italian speakers participated in a speeded lexical decision task in a fragment priming paradigm. Thirty bi- or tri-syllabic Italian nouns containing a biphonemic consonant cluster in internal position (e.g. *borsa*, 'bag') served as targets. Each target was primed by a sequence corresponding to an initial fragment of the target (e.g. *bor-borsa*). The fragment prime could consist of 3 o 4 phonemes and always ended with the first consonant of the cluster. The average length ratio between prime and target was 0.49. The clusters were different across words and each cluster could occur in only one target (although more than one fragment could end in a given consonant). 12 were heterosyllabic (e.g. *bor-sa* 'bag'), 12 tautosyllabic (e.g. *degrado* 'decay') and 6 ambisyllabic clusters (e.g. *dis-tanza* 'distance').

Another set of 30 Italian nouns matching for average length, frequency and prime/target length ratio, in which the fragment prime ended in a syllable onset consonant followed by a vowel (e.g. *tuc-tucano* 'toucan'). The same proportion of fragment-final consonants was maintained in the two sets of words.

Sixty pseudowords matching for average length and properties of the fragment were added. Pseudowords were obtained by changing one letter of existing words (belonging to the same frequency range of the experimental words), for

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

1/3 in their initial part, 1/3 in their central part and 1/3 in their final part. The 30 clusters used for pseudowords did not appear in the words' list.

In the lexical decision task, participants were asked to press a button corresponding to their dominant hand as soon as the orthographically presented target was judged as a word, and a different button for targets judged as nonwords. All the stimuli appeared in Courier New font, 18 point size in the center of the computer screen. In order to avoid allographic effects, primes were displayed in uppercase and targets in lowercase. The fixation was 200 ms, followed by a 50 ms pause. Primes appeared for 150 ms, followed by a 50 ms pause. The targets remained on the computer screen for a maximum of 1 sec. If the participants did not produce any answer within that time, the feedback *Fuori tempo* ('Out of time') appeared on the screen. Reaction times (RTs) and the number of errors (Nerr) constituted the dependent variables. The reaction times were measured from target onset to subject's response, and responses given after the deadline were scored as errors.

The Experiment was preceded by a practice session. When the participants reached the 70 % of valid responses the experiment started.

2.2 Experimental variables

Several statistical and distributional properties of word primes, targets and clusters were derived from the CoLFIS corpus (Bertinetto et al., 2005).

For each prime-target pair, we calculated (i) the token frequency of the target ('TargetFreq'), (ii) the N of words beginning with the prime fragment ('PrimeTypeFreq'), (iii) the cumulated frequency of the words in (ii) ('PrimeTokenFreq'), (iii) the length of the target (in N graphemes), (iv) the length of the prime (in N graphemes), (v) the prime/target length ratio.

For each cluster, we calculated (vi) the TP value, i.e. the probability with which the first consonant of the cluster predicts the occurrence of the following consonant, calculated over the corpus word tokens ('BigramTP'), (vii) the N of words containing the cluster ('BigramTypeFreq'), (viii) the cumulated frequency of the words in (vii) ('BigramTokenFreq'), (ix) the TP between the fragment prime and the second consonant of the cluster, e.g. P(*s*|*bor*) in *borsa* 'bag' ('SequenceTP'), (x) the N of words containing the sequence of the prime followed by the second C of the cluster ('SequenceTypeFreq'), (xi) the cumulated frequency of the words in (x) ('SequenceTokenFreq').

2.3 Analysis and results

Fixed and mixed models with subject and prime as random variables were used.

For the purposes of the present study, we tested two different models, both including frequency variables and phonotactic probability variables; they are shown in Table 1. The two models differed for the presence, in model II, of a measure of prime frequency, which was not included in model I, and for being focused either on sequence and bigram token frequencies (model I), or on sequence and bigram type frequencies. Both models were tested for CC items (e.g. *borsa*, 'bag') and CV items (e.g. *tuc-ano* ' toucan') separately.

	Model I	*Model II*
Fixed effects	TargetFreq LenghRatio SequenceTokenFreq BigramTokenFreq SequenceTP BigramTP	TargetFreq PrimeTokenFreq LengthRatio SequenceTypeFreq BigramTypeFreq SequenceTP BigramTP
Random effects	Subject Fragment prime	Subject Fragment prime

Table 1. Fixed and random effects for the CC and CV items.

The results of the fixed effects analyses for the relevant models are summarized in Table 2 (dependent variable: RTs) and Table 3 (dependent variable: Nerr).

According to model I, with RTs as the dependent variable, the sequence's TP (i.e., the TP between the fragment prime and the second consonant of cluster) turned out to be the most significant predictor, even outranking the contribution of frequency values (for the target, the sequence and the bigram), which all concurred to the intercept. A different picture emerged however for the CV items, for which no probability variables turned out to significantly predict the subjects' response times; on the contrary, the target frequency, with the secondary contribution of the frequency of the cluster, appeared to play a role for this subset of items.

According to model II, for CC items the role of the target frequency turned out to be very important, and the only additional effect was gener-

ated by the sequence's TP. Thus the two models were similar in emphasizing the role of the probability with which a given C follows the prime sequence. As for CV items, model II returned a picture very similar to the one that emerged in model I, with target frequency and bigram type frequency as the only significant predictors.

CC item model I					
	Estimate	Std.	Error	p-value	Adjusted R^2
Intercept	643.492	30.563	21.054	<0.001	0.433
TargetFreq	-12.008	5.062	-2.372	<0.05	
SequenceTokenFreq	-7.487	3.105	-2.412	<0.05	
BigramTokenFreq	-7.483	3.038	-2.463	<0.05	
SequenceTP	64.248	18.907	3.398	<0.01	
CV item model I					
	Estimate	Std.	Error	p-value	Adjusted R^2
Intercept	911.829	96.470	9.452	<0.001	0.47
TargetFreq	-15.849	3.860	-4.106	<0.001	
BigramTokenFreq	-24.888	8.513	-2.923	<0.01	
CC item model II					
	Estimate	Std.	Error	p-value	Adjusted R^2
Intercept	579.715	23.303	24.878	<0.001	0.3
TargetFreq	-17.809	4.867	-3.659	<0.01	
SequenceTP	43.518	18.859	2.308	<0.05	
CV item model II					
	Estimate	Std.	Error	p-value	Adjusted R^2
Intercept	838.777	90.123	9.307	<0.001	0.41
TargetFreq	-16.816	4.043	-4.160	<0.001	
BigramTypeFreq	-24.021	10.364	-2.318	<0.05	

Table 2. Fixed effects coefficients for the two models, CC and CV items (RTs=dependent variable).

When subject and prime were included as random factors, the pairwise comparison in the likelihood ratio test confirmed that the contribution of the sequence's TP increased significantly the predictability of the RTs patterns: $\chi^2(1)= 11.184$, $p= 0.0008$ in model I, $\chi^2 (1)= 5.4403$, $p= 0.019$ in model II.

The average reaction times and the number of errors were positively and significantly correlated, though with an intermediate correlation coefficient ($r = .648$, $p < .01$). We thus tested the two models with Nerr as the dependent variable, in order to determine if the error rate was influenced by frequencies and probabilities to a different extent than response latencies.

With Nerr as the dependent variable, R^2 values were consistently lower than in the RTs simulations (Table 3), thus indicating that the error patterns were accounted for by our frequency and probability variables to a more limited extent. In particular, both model I and model II emphasized for the CC items the role of target frequency as the only significant predictor of errors, while for CV items an additional role of bigram frequencies (by token and by type, respectively) was found. Thus for the CV items, RTs and error rate produced consistent results.

CC item model I					
	Estimate	Std.	Error	p-value	Adjusted R^2
Intercept	7.714	1.349	5.717	<0.001	0.2
TargetFreq	-1.207	0.462	2.613	<0.05	
CV item model I					
	Estimate	Std.	Error	p-value	Adjusted R^2
Intercept	61.5379	16.8337	3.656	<0.01	0.39
TargetFreq	-1.7089	0.6736	-2.537	<0.05	
BigramTokenFreq	-4.4807	1.4855	-3.016	<0.01	
CC item model II					
	Estimate	Std.	Error	p-value	Adjusted R^2
Intercept	7.714	1.349	5.717	<0.001	0.19
TargetFreq	-1.207	0.462	-2.613	<0.05	
CV item model II					
	Estimate	Std.	Error	p-value	Adjusted R^2
Intercept	55.4970	15.0637	3.684	<0.01	0.33
TargetFreq	-1.8956	0.6757	-2.805	<0.01	
BigramTypeFreq	-5.1472	1.7322	-2.971	<0.01	

Table 3. Fixed effects coefficients for the two models, CC and CV items (Nerr=dependent variable).

3 Discussion

This work aimed to shed light on the role of TPs in a so far unstudied experimental environment, i.e., a lexical decision task with fragment priming. As the large part of studies on phonotactic probabilities focused on English, this work also added to the field with evidence from a poorly investigated language, Italian.

Fragment priming is known to be modulated not only by word frequency and the frequencies of words matching the fragment but also by top-down information conveyed by the prime: a fragment prime matching a unique morpho-lexical family is as effective as a stem prime, thus showing that priming acts as a cue for the properties displayed in the target (see e.g. Laudanna & Bracco, 2006, for Italian).

This study has shown that the priming effect when an initial fragment is available is influenced also by bottom-up variables; in particular, it depends on the probability with which the segments composing the fragment or the fragment-final consonant predict the occurrence of the consecutive consonant. Although to a lesser extent, the frequency with which bigrams and sequences occur (as types or tokens) in the lexicon also predict the subjects' behavior. Phonotactic probabilities thus turned out to predict the subjects' response to a large degree for many of the phonological environments tested in the current experiment, sometimes outperforming target frequencies, and consistently overtaking the contribution of the prime/target length ratio and of the prime frequency.

The results however suggested that the phonotactic probabilities in the case of consonant clusters were overall more important than in the case of consonant-vowel sequences; thus it must be

concluded that the contribution of TPs in lexical recognition is not the same across phonological environments. Consonant clusters might play a particularly relevant role in lexical access, compared to CV sequences, as contemporary theories based on the principles of phonological and morphological naturalness also seems to predict (see e.g. Dressler & Dziubalska-Kolaczyk, 2006; Korecky-Kroell et al. 2014).

Additionally, for CC sequence the token frequencies (of the bigram and of the prime + C sequence) turned out to be relatively more important than the corresponding type frequencies, thus suggesting that the exposure to the number of occurrence of a cluster or of a segment sequence may be more important in lexical access than the exposure to the individual items containing them.

An additional issue concerns the role of TPs in morphologically complex words. According to some models, morphological parsing is necessary for lexical access and the prefix (in the case of prefixed words) has to be stripped away in order for the word to be recognized (from Taft & Forster, 1975 onwards). Assuming a condition in which the fragment prime coincides with a prefix, TPs would play the additional role of marking the morphological boundary during the priming event. According to the results of the current study, it appears to be of utmost importance to further verify whether prefixed and pseudo-prefixed words behave in the same way. In fact, models postulating morphologicl pre-parsing (e.g. Schreuder & Baayen, 1995) would suggest that high TPs will codetermine latencies for prefixed targets only, while if morphology does not affect word recognition, then the TPs between the fragment prime and the following segment composing the target will modulate latencies in prefixed and pseudo-prefixed words to the same extent.

A follow-up experiment will therefore test the contribution of phonotactic statistical knowledge in native speakers' access to complex word forms (specifically, prefixed nouns). Prefixed and pseudo-prefixed words will be used for that purpose. In particular, fragment primes will be selected according to two different conditions: in condition a) the targets are prefixed words and the fragment prime coincides with the prefix (e.g. *bis-bisnonna* 'grandmother'); in condition b) the targets are pseudo-prefixed words and no morphological boundary occurs between the initial fragment and the second part of the word (e.g. *per-perdente* 'loser'). Together with the current experiment, the experiment on prefixed and pseudo-prefixed words will determine whether or not the role of TPs is different when the target is a simplex word compared to when it is a prefixed word, and to when it is a pseudo-prefixed word. Different hypotheses may be put forward here, according to whether or not morphological boundaries affect the processing of consonant clusters (e.g., Calderone et al. 2014, Celata et al. 2015 in press), and according to the likelihood that a given sequence occurs as morpheme or as homographic non-morphological pattern (see Laudanna et al., 1994).

By describing phonotactic probability and frequency effects during word recognition, this study offers arguments to models of lexical access based on bottom-up processes such as cohort models for orthographic stimuli (see e.g. Johson & Pugh, 1994). The property of single consonants to predict the following segment then speeding up the recognition of the whole word, as an additional if not independent way to access words and their subparts, might also be discussed with reference to models that associate orthographic input units to semantic and lexical knowledge (from connectionist models such as in Harm & Seidenberg, 1999, to amorphous models such as in Baayen et al. 2011).

References

Harald R. Baayen. 2010. A real experiment is a factorial experiment? *The Mental Lexicon*, 5(1): 149-157.

Harald R. Baayen, Petar Milin, Dusica Filipovic Durdevic, Peter Hendrix and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension on naive discriminative learning. *Psychological Review*, 118: 438-482.

Pier Marco Bertinetto, Cristina Burani, Alessandro Laudanna, Lucia Marconi, Daniela Ratti, C. Rolando and Anna Maria Thornton. 2005. *Corpus e Lessico di Frequenza dell'Italiano Scritto CoLFIS)*. http://linguistica.sns.it/CoLFIS/Home.net

Basilio Calderone, Chiara Celata, Katharina Korecky-Kroell and Wolfgang U. Dressler. 2014. A computational approach to (mor)phonotactics: Evidence from German. *Language Sciences*, 46 (part A): 59-70.

Chiara Celata, Katharina Korecky-Kroell, Irene Ricci, and Wolfgang U. Dressler. 2015 (in press). Online processing of German (mor)phonotactic clusters by

adults and adolescents. *Italian Journal of Linguistics*, 27(1).

Wolfgang U. Dressler and Katarzyna Dziubalska-Kolacyk. 2006. Proposing Morphonotactics. *Italian Journal of Linguistics*, 18: 249-266.

Katharina Korecky-Kroell, Wolfgang U. Dressler, Eva Maria Freiberger, Eva Reinisch, Karlheinz Moerth and Gary Libben. 2014. Phonotactic and morphonotactic processing in German-speaking adults. *Language Sciences*, 46 (part A): 48-58.

N.F. Johnson and K.R. Pugh. 1994. A cohort model of visual word recognition. *Cognitive Psychology*, 26: 240-346.

Alessandro Laudanna, Cristina Burani and Antonella Cermele. 1994. Prefixes as processing units. *Language and Cognitive Processes*, 9, 295-316.

Alessandro Laudanna and Giulia Bracco. 2006. Stem and fragment priming on verbal forms of Italian. *In Proceedings of the 5th International Conference on the Mental Lexicon* (Montreal, Canada, 11-13 October, 2006): 26.

Paul A. Luce and Nathan R. Large. 2001. Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, 16: 565-581.

Sven L. Mattys and Peter W. Jusczyk. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78: 91-121.

Mark Pitt and James McQueen. 1998. Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, 39: 347-370.

Robert Schreuder and Harald R. Baayen. 1997. How simplex complex words can be. *Journal of Memory and Language*, 37: 118-139.

Holly L. Storkel. 2001. Learning nonwords: Phonotactic probabilities in language development. *Journal of Speech, Language, and Hearing Research*, 44: 1321–1337

Marcus Taft and Kenneth I. Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14: 638-647.

Michael Vitevitch, Paul Luce, J. Charles-Luce and D. Kemmerer. 1997. Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, 40: 47–62.

Michael S. Vitevitch and Paul A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language*, 40: 374-408.

Michael Vitevitch, Paul A. Luce, David B. Pisoni and Edward T. Auer. 1999. Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language*, 68: 306-311.