Identity Avoidance Between Non-Adjacent Consonants in Artificial Language Segmentation

Natalie Boll-Avetisyan & René Kager

Phonotactic distributions over lexicons correlate with human performance (Coleman & Pierrehumbert, 1997; Vitevitch & Luce, 1999; Pitt & McQueen, 1998). Although it is known that phonotactic probability is independent of lexical analogy (Bailey & Hahn, 2001), the issue remains whether gradient distributions over the lexicon are represented in the grammar in the form of abstract constraints involving natural classes (Frisch, Pierrehumbert, & Broe, 2004; Hayes & Wilson, 2007).

This paper supports the hypothesis that gradient distributions over the lexicon may be represented as abstract constraints, which affect segmentation. We tested this hypothesis using artificial language segmentation, a task which is known to be affected by phonological properties of the native language (Onnis, Monaghan, Chater, & Richmond, 2005). We focussed on non-adjacent C_C dependencies, allowing us to test effects of abstract constraints while controlling for phonotactic probabilities of the stimulus material. It has been found that statistical calculations over triphones (e.g. CVC) are ineffective for non-word processing (Bailey & Hahn, 2001). Yet artificial language learning studies show that statistical dependencies between non-adjacent consonants are learnable (Newport & Aslin, 2004). Many languages restrict the co-occurrence of homorganic consonants across intervening vowels (e.g. Frisch & Zawaydeh 2001). Artificial language studies and cross-linguistic studies are consistent with the prediction that constraints on non-adjacent consonants with a probabilistic basis in the lexicon affect artificial language segmentation.

Lexical statistics

We collected cooccurrence statistics on non-adjacent consonants in Dutch using Observed/Expected values, with E(C1VC2) calculated as p(C1) * p(C2) * N_CVC in CELEX (Baaijen, Piepenbrock, & Gulikers 1995). Two observations stood out. Pairs of labials (P={p b f v m w}) are strongly under-represented (O/E = 0.45) as compared to coronal pairs (T={t d s z n l r}; O/E = 0.77). In word-initial position (where E = p(C1) * N_C), labials are strongly over-represented (O/E = 2.0), whereas coronals are only slightly under-represented (O/E = 0.60). From this distribution, we hypothesized two abstract constraints. OCP-LAB forbids pairs of consecutive labials. ALIGN-LAB requires labials to occur word-initially. In order to examine whether these constraints are used for segmentation, we ran two artificial language experiments.

Experiment 1

Native speakers of Dutch were trained in an artificial language that consisted of six P syllables (P1={p o, be, ma}, P2={pa, bi mo}) and three T syllables (T={tu, do, ne}), which were concatenated into a synthetic speech stream (…P1P2TP1P2TP1P2T….) with flat prosody. Other potential cues for segmentation from lexical statistics, such as transitional probabilities between syllables and segments, positional syllable frequency, and the distribution of nasality and voicing were controlled. If participants were able to segment words from this stream, predictions from OCP-LAB and ALIGN-LAB were that that PTP words should come out best, and TPP words should come out worst.

In the test phase, participants had to decide which of two CVCVCV strings was a word from the language they had been trained on. Participants (N=42) were assigned to one of three test conditions: PTP-PPT, PTP-TPP, or PPT-TPP. The 48 test pairs were matched for lexical factors (cohort density, lexical neighbourhood density). As predicted, PTP words, which satisfy OCP-LAB and ALIGN-LAB, were preferred over PPT (p < .001) and TP (p < .05) to be words of the artificial language in which participants had been trained (see Figure 1). In particular, the strong preference PTP > PPT suggests that OCP-LAB has an impact on segmentation. Results of the individual test items did not correlate with O/E values for consonant combinations, which supports our hypothesis that segmentation preferences reflect an abstract constraint,
rather than experience with individual consonant combinations. In the PPT-TPP comparison, the predicted preference for PPT, due to ALIGN-LAB alone, was not found. This suggests that the PTP segmentation may have been too dominant over the PPT or TPP segmentations for a PPT-TPP comparison to be possible at all. A replication of the experiment using a different syllable set yielded the same result.

**Experiment 2**

In order to rule out a possible interpretation that the result was not caused by OCP-LAB, but by a general perceptual preference for identity at edges (ABA > AAB, ABB, regardless of values of A,B), we ran a second experiment. Participants were trained in an artificial language similar to the language in Experiment 1, the only difference being that P1 was replaced by T (T2={ta, di, no}). In Dutch, consecutive coronals are less under-represented than labials. Accordingly, we hypothesized that any co-occurrence effects will be over-ruled by ALIGN-LAB, and expected a PTT segmentation.

As in Experiment 1, we assigned participants (N=42) to one of three test conditions, contrasting TPT-TTP, TPT-PTT, or PTT-TTP pairs. As predicted, PTT words were favoured over TPT (p < .001) and TTP (p < .001). In the TTP-TPT comparison, none of the words was favoured, which supports the prediction that ALIGN-LAB affected segmentation, not a general preference for ABA.

**Conclusion**

The segmentation results suggest that probabilistic distributions over the lexicon may be represented as abstract constraints in the grammar. The hypothesized constraints were abstract since their effects could not be reduced to low-level statistical properties of the stimulus material. Future studies may explore the consequences of our findings for models of speech segmentation.

**References**


