To appear in Laboratory Phonology: Journal of the Association for Laboratory Phonology
Please cite as:

Kevin Tang & Faruk Akkus. 2023 Identity avoidance in Turkish partialreduplication: feature specificity and locality. Laboratory Phonology: Journal of the Association for Laboratory Phonology 14(1). doi:10.16995/labphon.6459

Identity Avoidance in Turkish Partial Reduplication: Feature Specificity and Locality

Kevin Tang[1] and Faruk Akkuş[2]

[1]Department of English Language and Linguistics, Institute of English and American Studies, Heinrich-Heine-University, Düsseldorf, 40225, Germany
[1]Department of Linguistics, University of Florida, Gainesville, Florida, 32611-5454, United States of America
[2]Department of Linguistics, University of Massachusetts Amherst, MA, U.S.A., 01002

Corresponding author: Kevin Tang (kevin.tang@hhu.de)

Abstract

This study investigates the Turkish partial reduplication phenomenon, in which the reduplicant is derived by prefixing C_1VC_2 syllable, where C_1V are identical to the word-initial CV of the base and the C_2 ends in one of the four linking consonants: *-p, -m, -s, -r*. This study re-examines the factors conditioning the choice of the linking consonant, by focusing the nature of the (dis)similarity (feature specificity) and the proximity (locality) between the consonants in the base and the linking consonant, using an acceptability rating task with over 200 participants and a diverse set of stimuli in terms of length and word shapes. Results indicate a gradient identity avoidance effect that extends over all consonants in the base. Crucially, the effect of all consonants is not uniform, with the strength of the effect decreasing further into the base. The study also uncovers an elusive interplay between the distance-based decay effect and the syllable position effect, both of which turn out to play a role in these non-categorical patterns with multiple features. Furthermore, results indicate that identity avoidance operates over both individual features as well as whole segments. Overall, the study argues that locality-sensitive feature-specific identity avoidance constraints are part of the grammar.

Keywords: reduplication, locality, similarity, feature specificity, syllable role, OCP, Turkish

1. Introduction

Many languages exhibit a process in which a consonant is altered in order to reduce its resemblance partially or fully to another consonant (see e.g., Stanton 2017; Suzuki 1998 for some discussion of (long-distance) consonant dissimilation). A representative example of dissimilation comes from Latin, in which certain suffixes exhibit dissimilatory effects with respect to the consonants in the stem. For example, when the stem does not contain any lateral consonant [l], the adjectival suffix *-alis* surfaces as is (1a-b). However, when the stem contains a lateral consonant at any position in the stem constituent, the suffix surfaces as [-aris] (1c-d).

(1)	a.	/nav-alis/	nav-alis	'naval'
	b.	/episcop-alis/	episcop-alis	'episcopal'
	c.	/sol-alis/	sol-a**r**is	'solar'
	d.	/lun-alis/	lun-a**r**is	'lunar' (Suzuki 1998:12,(3))

In this regard, reduplication is a particularly fruitful process to examine since although in many cases the material attached to the base in reduplication resembles the base phonologically, exact copying of the base is not always achieved (see e.g., Alderete et al. 1999; Frampton 2009 for some illustrations). As such, similarity between the reduplicated form and the (relevant part of the) base is reduced. Against this background, we analyze a pattern of reduplication from Turkish with a partial dissimilation, which turns out to be informative as to several open questions related to the effects of the identity avoidance,[1] such as at what granularity it applies (e.g., at the level of segments or phonological features), the extent to which proximity (locality) plays a role, or its interaction with syllable roles, especially in instances with multiple features involved.

Cross-linguistic studies have found that features may differ in their strength in phonological patterns that involve identity avoidance (see an extensive survey of 46 phenomena by Bye (2011)). While many feature classes have been found to participate in identity avoidance, such as place of articulation, laryngeal state, manner (continuancy, liquid, nasality), vowel height, and suprasegmental properties (length and tone), these typological analyses are based mostly on patterns that involve only one or two featural changes. Analysing such patterns means that the researchers are able to clearly identify the relevant features. For example, only the [lateral] feature participates in identity avoidance in the Latin liquid dissimilation. Looking beyond phonological patterns, several studies (Berkley, 2000; Frisch, Pierrehumbert, & Broe, 2004; Graff & Jaeger, 2009) have examined the statistical asymmetries in the lexicon which suggests that there is preference for phonetically dissimilar consonants within a word, such that fewer similar sounds co-occur than would be expected *a priori*. This suggests that identity avoidance plays a role in shaping the phonological lexicon. This type of analyses provides a fruitful ground for better understanding the typology of participating features in identity avoidance. This is because one must examine a larger set of features simultaneously, owning to the logically possible combinations of consonants in a root, as opposed to just one or two features that participate in a phonological pattern. However, there exists another type of phonological phenomena that involve identity avoidance with a larger set of phonological features, such is the case of Turkish partial reduplication, which involves four consonants, potentially allowing multiple features to participate in dissimilation. Looking at these phenomena would further inform our understanding of the typology of identity avoidance.

The partial reduplication in Turkish also helps further our understanding of *syllable position* in assimilation (and dissimilation), a phenomenon examined in Bennett (2012, 2013); Rose and Walker (2004); Suzuki (1998). Focusing on R-dissimilation and L-assimilation in Sundanese as well as nasal assimilation (i.e., nasal agreement/harmony) in Kikongo, these studies find the following patterns, which are also categorical as they usually involve one or two phonological features: R-dissimilation occurs whenever the affix /r/ and an /r/ in the root have different syllable roles. In the case of both L-assimilation and nasal assimilation, however, matching syllable roles contribute to segments' similarity allowing assimilation to take place between segments that share the same syllable role. For example, in the case of Sundanese L-assimilation, the initial /l/ of the root and the /r/ of the affix /ar/ end up as the onsets of the first two syllables of the stem. In fact, these (mis)match patterns

[1] In this paper, we use 'identity avoidance effect' for the most part, instead of the Obligatory Contour Principle (OCP) as a more neutral term although we still refer to OCP as well occasionally.

are given as predictions in Bennett (2013, p. 536) using the constraint CC·SRole, which limits correspondence based on structural position. Accordingly, Bennett argues that while harmony/assimilation is predicted between consonants with matching syllable roles, dissimilation is predicted for consonants with mismatching syllable roles (but not those with matching syllable roles). Various questions arise in light of these patterns: first, is it possible to find dissimilation that would favor matching syllable roles? Secondly, can the effect of syllable role be still observed in non-categorical patterns with multiple features, and relatedly, what are the effects (if any) of syllable role in gradient assimilation/dissimilation?

Moreover, little is known about the interaction of the effect of syllable role with the *distance-decay effect*. Zymet (2014) illustrated that in multiple languages the likelihood of application of assimilation or dissimilation decreases as transparent distance increases. The patterns that were examined involve usually one or two phonological features, such as rounding dissimilation in Malagasy, liquid dissimilation in Latin and in English, and vowel harmony in Hungarian. It remains an open question whether distance-based decay can still be observed in patterns that involve with multiple features, and whether there is an interplay between the distance-based decay and syllable position.

Turkish is apt to address – at least provide significant insights to – these open questions because the partial reduplication in Turkish is a gradient dissimilation phenomenon, which involves four distinct consonants, potentially allowing multiple features to participate in dissimilation. Furthermore, it shows a strong effect in matching syllable roles, and exhibits an interesting interplay between the role of syllable position and the distance-based decay.

As will be shown below, Turkish has a type of emphatic reduplication, in which the emphatic variants are derived by prefixing a C_1VC_2 syllable, where C_1V are identical to the word-initial CV of the base, and the C_2 (known as *linking consonant)* ends in one of the four consonants (*-p, -m, -s, -r*), effectively reducing the resemblance with the base. The effect of consonant dissimilation has largely been attributed to the first two consonants of the base (Kelepir 1999, 2000; Wedel 1999). This study examines the nature of the factors behind the choice of the C_2 using an acceptability rating task with over 200 participants and a more diverse set of stimuli in terms of length and word shapes than the ones employed in prior literature. While partially confirming the conclusions of some prior studies, the current study reveals novel findings. The effect beyond the first two consonants have been mentioned but they are not often found to be statistically significant or required when the first two consonants have already been considered (e.g., Yu 1999). This pattern suggests that locality in which dissimilation operates in a categorical manner. This is particularly significant in light of the recent discussions of distance-based decay effect (Zymet 2014, 2018), which, as mentioned above, suggests that the likelihood for the application of a phonological process decreases as the distance increases, and this can happen in a gradient and non-linear manner. This study establishes that the effect of identity avoidance spreads across all consonants in the base, which crucially is not uniform but exhibits a distance-based decay effect. This effect also interacts with the syllable position effects. In particular, the effect would decay with the distance from the linking consonant and would be enhanced if the consonant in the base matches the constituency of the linking consonant (coda vs onset). For instance, in the case of Turkish, the linking consonant is always a coda (in the C-initial forms, which is the focus of this study). When a base form has the word shape C_1VC_2V, a distance-based decay effect was found, as hypothesised, with C_2 being less influential than C_1. However, the word shape C_1VC_2 exhibits the opposite pattern with C_1 being less influential than C_2. This surprising effect can be attributed to that the LC and the C_2 in C_1VC_2 are both codas. Accordingly, the Turkish phenomenon is particularly insightful in this regard too: it showcases the interplay of distance-based decay effect and syllable position effect, which in most studies are treated separately (due to the data involved). Additionally, in line with findings in cross-linguistic typology, features do not participate equally in identity avoidance processes, with some being more influential than others. Our study confirms the importance of only some of the features employed in previous studies, such as [strident], [labial] and [nasal], but not others, such as [coronal], [sonorant], or [continuant].

These new findings contribute to our understanding of the nature of locality, particularly arguing for the view that locality-sensitive feature-specific identity avoidance constraints are part of the grammar. This study overall highlights the value of revisiting a long-debated topic with a new lens to address any remaining unsolved puzzles, replicate existing findings, and generate new hypotheses that can contribute to the future studies.

The rest of the Introduction section introduces the phenomenon itself, and situates its implications within the larger literature with a focus on locality/syllable role and feature-specificity.

1.1. The phenomenon and the implications

Turkish is an agglutinative language, with the majority of derivation achieved through suffixation. A rare instance where prefixation is observed in the language is the so-called *partial reduplication* (or *emphatic reduplication*) (Demircan 1987; Lewis 1967, i.a.). It is the only prefix present in the language except for foreign prefixes in borrowed forms. The partial reduplication is found with modifiers, namely adverbs and adjectives. As shown in (2)[2], emphatic variants are derived by prefixing a C_1VC_2 syllable. The initial C_1V are identical to the word-initial CV of the base. However, C_2 ends in one of the four consonants: *-p, -m, -s, -r* (Lewis 1967), generally referred to as *linking consonants (LC)*.[3]

(2)	**Base**	**Gloss**	**Reduplication**	**Gloss**
	kara	'black'	*ka**p**-kara*	'very black'
	beyaz	'white'	*be**m**-beyaz*	'very white'
	ma:vi	'blue'	*ma**s**-ma:vi*	'fully blue'
	temiz	'clean'	*te**r**-temiz*	'completely clean'

On the other hand, if the base is V-initial, the LC surfaces as *-p*.

(3)	**Base**	**Gloss**	**Reduplication**	**Gloss**
	açık	'open'	*a**p**-açık*	'very open'
	ince	'thin'	*i**p**-ince*	'very thin'
	ansızın	'suddenly'	*a**p**-ansızın*	'very suddenly'

Partial reduplication in Turkish has been the subject of many studies, which have aimed to explain the conditions underlying the choice of the LCs (see e.g., Demir 2018; Demircan 1987; Dobrovolsky 1987; Foster 1969; Hatiboğlu 1973; Kaufman 2014; Köylü 2020; Lewis 1967; Sofu 2005; Sofu and Altan 2008; Taneri 1990; Wedel 1999; Yavaş 1980; Yu 1998, 1999). This phenomenon is particularly suited for shedding light on the nature of the identity avoidance effect in terms of its locality and feature specificity. It is a type of dissimilation in reduplication with *fixed segmentism* (Alderete et al. 1999). Crucially, prototypical instances of reduplication with fixed segmentism involve only a single fixed unit, e.g., [a] in Javanese habitual reduplication (Yip 1997:18), or [m] in Turkish, (4), in which an invariant segment appears whatever the features of the base. However, as seen in (2), the

[2] In the transcription, the following correspondences hold between the orthographic forms and IPA: *ı* → [ɯ], *ç* → [t͡ʃ], *c* → [d͡ʒ], *ş* → [ʃ], *ü* → [y]

[3] In terms of the morphological analysis, we assume that the reduplication is RED + LINKER + BASE where RED is only CV and only the LINKER can dissimilate.

number of fixed segments is four in Turkish emphatic reduplication. This allows for many more features to participate in the identity avoidance effect.

(4)	**Base**	**Gloss**	**Reduplication**	**Gloss**
	sarı	'yellow'	*sarı **m**arı*	'yellow or similar colors'
	kapı	'door'	*kapı **m**apı*	'door or the like' (Turkish)

Another interesting property of the Turkish emphatic reduplication is that, it differs from most crosslinguistic examples in not just allowing multiple possible fixed segments, but also having variability in which LC would appear for a given stem, as exemplified in (5). As such, although it is not always reported in the literature (though see Müller 2003; Wedel 1999; Yu 1999), there is indeed variability with respect to the choice of the LC with a substantial number of bases.

(5)	**Base**	**Gloss**	**Reduplication**	**Gloss**
	yeşil	'green'	*ye**m**/**p**-yeşil*	'completely green'
	başka	'different'	*ba**m**/**p**-başka*	'very different'
	buruşuk	'creased'	*bu**m**/**s**-buruşuk*	'very creased'
	yırtık	'torn'	*yı**s**/**p**-yırtık*	'completely torn'

Moreover, with items that allow multiple LCs, variation exists as to which LC is preferred for a given item across speakers. Even the existence of a varying degree of preference for the choice of a particular linking consonant (and between multiple LCs) differentiates the Turkish phenomenon from the classic reduplication with fixed-segmentism. Crucially, in Turkish partial reduplication, as we will argue in this paper, the choice of the linking consonant is sensitive to the features of the base. In some instances of reduplication with fixed segmentism (e.g., Alderete et al. 1999; McCarthy and Prince 1993), the identity of that fixed segment is attributed to the emergence of the unmarked, default form. However, as has been noted by Wedel (1999) and Yu (1999), this is not the case in Turkish, and these LCs appear in the output despite clear markedness violations and are not the unmarked segments in the language. The consonants [p, m, s] are not considered the default segments in Turkish, and instead [n, j] are considered to be so (Wedel, 1999). As such, the latter are known as 'buffer consonants' in traditional Turkish grammars, appearing in various contexts e.g., breaking up vowel hiatus as in *kedi-**y**-i* 'cat-[j]-Accusative' "the cat".

While the seminal work by Demircan (1987) and subsequent work all identified the identity avoidance effect as a major factor, their accounts often involve heuristically chosen features that operate over the first two consonants in the base.[4] The current study (re)-examines the factors conditioning the choice of the LC, focusing on the C-initial forms since the V-initial forms are consistently reduplicated with the LC *-p*. We extend the previous studies by examining the nature of the proximity (locality) and the (dis)similarity (feature specificity) between the consonants in the base and the LC. Starting with the former, we examine both topics in turn, particularly how they have been handled in the Turkish literature on emphatic reduplication as well as their implications for the broader literature beyond Turkish.

[4]While the heuristic choice of the features might be a well-established way of doing traditional phonological analysis (Kenstowicz & Kisseberth, 2014, Ch. 2), which we also appreciate, this study demonstrates that a more statistical approach that does not rely on heuristic choices reveals properties that would otherwise be missed. Thus, our goal is not to critique the use of heuristics per se, but to highlight the point that in some cases we can learn more from a statistical approach.

1.2. Locality avoidance in Turkish and beyond

With respect to locality, most studies emphasise the importance of C_1 and C_2, yet some of them (implicitly) assume a cut-off after C_2. One of the questions this study addresses is whether the consonants beyond C_2 do not play a role in the choice of the LC. Let us begin by taking a closer look at the classic study on the phenomenon by Demircan (1987) which observes that the selection of the linker is subject to various dissimilation constraints. The primary observations are given in (6), in the format they are succinctly summarized in Yu (1999, p.5 & p.18) and with slight modifications.[5]

(6) Demircan's (1987) observations (adapted from Yu 1999)

- (i) The linker cannot be identical with any of the consonants in the base.
- (ii) No gemination: The linker should not be identical to the C_1 of the base.
- (iii) Avoid full reduplication: The linker cannot be identical with C_2 of the base with C_1VC_2 items.
- (iv) Featural identity avoidance: Avoid a linker that shares similar features, such as [labial], [strident] & [approximant], with any segment in the base.

Note that the observations (ii) and (iii) by Demircan focus on the contrast between the linking consonant and the C_1 and C_2 of the base, while the observations also make reference to the whole base. However, this latter point has been underappreciated in some subsequent studies. For example, based on the minimal pair in (7), Kelepir (1999, 2000) argues that not only C_1 of the base, but also C_2 matters for the choice of the LC.

(7)	a.	*yeni* 'new'	***yep**-yeni* 'completely new'
	b.	*yeşil* 'green'	***yem**-yeşil* 'completely green' (Kelepir 2000:11)

Similar to Demircan's (1987) observation (i), Kelepir's (1999) study also has a constraint/restriction, *Repeat [strident], which makes reference to the whole base (specifically, that rules out the strident linker [s] if there is a strident in the base). Yet, Kelepir's constraint system centers around the comparison of the LC with respect to C_1 and C_2 of the base.

Wedel (1999, 2000) is another study that argues for the presence of dissimilatory phonological constraints in partial reduplication, yet lacks a constraint of the sort proposed by Demircan and Kelepir, who allowed for the scanning of the whole base to avoid the choice of an LC that shares all or some of the features of the LC. In fact, the author explicitly mentions that there should be a cut-off after C_2. This is reflected in the generalizations and the OT constraints that Wedel proposes. Consider (8):

(8)
- (i) [p] is not selected if C1 is labial: *PLOSIVE-αPL
- (ii) The interpolated consonant [LC] must be non-identical to C1: *GEM
- (iii) The interpolated consonant [LC] must be non-identical to C2: *REPEAT (Wedel 2000: 550)

The assumption that consonants beyond C_2 have no significant effect has led some researchers that do nonce-word studies to only create nonce-words consisting of two-consonant

[5]The summaries are Yu's recasting of Demircan's observations about dissimilation constraints, thus has some slight modifications which do not make a difference for the content. For example, Yu (1999) uses the term *linker*, while Demircan 1987 uses the term 'closer' for the LC. The observation (i) is given as "Avoid closers identical with any of the base consonants to rule out..." in Demircan 1987, while it is given as "The linker cannot be identical with the final consonant of the base" in Yu 1999 with some other qualifications.

stems. This is most clearly seen in the study of Köylü (2020) whose all 48 nonce-words consists of bases with a maximum of two consonants.

In fact, examining the features labial, coronal and strident, Yu (1999) found that the strident feature has an effect with respect to C_3, and corroborates Kelepir's *Repeat [strident] constraint. More generally, this supports the observation that the LC is not restricted to identity avoidance effect with respect to C_1 and C_2. Crucially, Yu did not find a significant effect from segments beyond C_3 which he attributed to the limited number of adjectives and adverbs with more than three consonants in his corpus data.

To sum up, the role of C_1 and C_2 has been the focus of many previous studies of the Turkish phenomenon.[6] There is nonetheless evidence from other languages that suggests all the consonants in most dissimilation phenomena could play a role in identity avoidance, being also subject to a *distance-based decay effect* (Zymet, 2014, 2018). This effect states that the likelihood for the application of a phonological process decreases as transparent distance increases. Arabic is the poster child for this kind effect in identity avoidance (see e.g., Coetzee and Pater 2008; Frisch et al. 2004; Frisch and Zawaydeh 2001; McCarthy 1986; McCarthy and Prince 1994). For example, OCP-Place is a single gradient constraint that restricts consonant co-occurrence in Arabic based on (i) similarity (see below for the discussion of similarity) and (ii) proximity (Frisch et al., 2004; Frisch & Zawaydeh, 2001). Particularly, the influence of similarity on consonant co-occurrence is affected by distance, as the constraint is weaker for non-adjacent consonants. Accordingly, the fact that many prior studies on Turkish partial reduplication consider only C_1 and C_2 and the null effect of segments beyond C_3 in Yu (1999) might also be due to the distance-based decay effect. The present study indeed uncovers that the identity avoidance in Turkish is subject to such an effect, where it holds for all the segments, with the effect being strongest from C_1 and being weakened as a function of its distance from the target segment.

1.3. (Dis)similarity of features in Turkish and beyond

The (dis)similarity of features plays an even more prominent role in the identity avoidance literature. For example, studies usually aim to probe which specific features identity avoidance constraints are sensitive to (whether they are categorical, Bennett 2013, or gradient, see e.g., Frisch et al. 2004; Gallagher and Coon 2009; McCarthy 1986), or whether features matter to the same degree. If the answer to the latter is no, then what is the extent to which specific features matter as opposed to other features in a given language and cross-linguistically? (see e.g., Bye 2011; Coetzee and Pater 2008; Frisch et al. 2004; Gallagher and Coon 2009; Graff and Jaeger 2009) It turns out, typologically, not all features participate equally in identity avoidance. Based on an extensive survey of 46 phenomena by Bye (2011), the following phonological dimensions have been found to play a role: the place of articulation, the laryngeal state, the manner of articulation (continuancy, liquid, nasality), vowel height, and suprasegmental properties such as length and tone. However, major class features such as [consonantal], [sonorant] and [approximant] do not play much of a role. Concerning the place feature, [labial] is relatively common, while [coronal] is rare and [dorsal] is unattested. Furthermore, alternations involving laterals and rhotics are relatively common.

When we look at previous studies on Turkish, we observe that they tend to focus on specific (set of) features. For instance, the analysis in Kelepir (1999) is built on avoidance constraints that use specific features. In particular, base consonants contrast with their correspondents in the reduplicant in place and sonorancy, i.e., the features [coronal], [sonorant], [labial], or [continuant]. Demircan (1987) identifies [coronal], [labial] and [nasal]

[6] But see Demircan 1987; Yu 1999.

as features of importance, and Yu (1999) adds that [strident] and [approximant] features also play a role in the interaction of segments.

Some remarks are in order regarding these previous studies on Turkish and their connection to the larger literature. Firstly, the choice of the relevant features is usually heuristic. For instance, Wedel (1999, 2000) uses PLOSIVE-αPL to explain why [p] is not selected if C_1 in the base is labial, while Kelepir (2000) uses the exact consonant [b] directly through the constraint, *-pb-. The heuristic nature mostly stems from the fact that researchers posit features, OT-constraints or perception-related restrictions that can explain the respective datasets. In fact, a lot of the proposed constraints are correlated and refer to overlapping issues, in that a constraint might be implicated or even entailed by another constraint. For example, the rule (i) in (6) encompasses the rule (ii). Secondly, and relatedly, it is unclear whether these features are the only features that matter, or perhaps there are some other features that may also trigger identity avoidance, but have been missed since most studies rely on the researcher's informed observations rather than a systematic examination of all possible features. Thirdly, cross-linguistic studies have found that features may differ in their strength in identity avoidance. For example, Gallagher and Coon (2009) find that [+strident] and [+ejective] have greater effects than others in Chol, while non-coronal place features were found to play a greater role in Arabic and Muna (Coetzee & Pater, 2008; Pierrehumbert, 1993). As such, various features have been shown to trigger the identity avoidance effects of varying strengths in different languages. Several proposals have been put forth with the aim of capturing the varying strength of the features. For example, Frisch et al. (2004) proposes a *similarity metric of natural classes*, which are shown to strongly correlate the observed-over-expected ratios (O/E) of consonant pairs co-occurring within Arabic roots, in a way that relies on the feature inventory of a language. However, data from other languages have not been able to replicate the effect of this metric. Coetzee and Pater (2008) failed to establish a correlation using this metric for Muna and Rotuman. Their discussion of place of articulation patterns in Muna and Arabic also demonstrates that the relative strength of place co-occurrence patterns cannot be due to inventory structure alone. Similarly, Graff and Jaeger (2009) found no evidence for such a correlation using this metric by any of the three languages examined (Dutch, Aymara and Javanese). An alternative approach was proposed by Graff and Jaeger (2009) which did find effects of feature-specific effects in identity avoidance of the three language using a more complex model. This model allows each individual feature from each consonant in the base being weighted freely.

Additionally, previous analyses involved using a specific set of features, but their formulation of these features are often correlated, for example, a constraint is entailed by another. The choice of relevant features, therefore, tends to be heuristic. This is partly due to feature redundancy, a design property of standard versions of feature theory. Multiple sets of relevant features that could play a role in consonant dissimilation. For example, one could use any of following features [nasal], [labial], [sonorant], [voice] to model the dissimilation with /m/. That is not to say redundant feature values do not matter. In fact, Keyser and Stevens (2006); Stevens and Keyser (1989) have proposed that redundant feature values can enhance the phonetic interpretation of contrastive values (see Clements and Ridouane (2006) for an overview). On top of feature identity avoidance, previous analyses also make reference to total identity of C_2, such as C_2 cannot be identical to any of the consonants in the base. Again, these two types of factors are correlated. Previous analyses have also not fully addressed the question of whether the effect of total identity can be reduced to the effect of feature identity. These considerations call for revisiting the phenomenon with a different theoretical angle[7] as well as new methodological tools that can jointly evalu-

[7]For example, this study is not concerned with the particular angle some prior studies take, i.e., whether the Turkish partial reduplication fits into a phonological analysis (Emergence of the Unmarked), or morphological

ate identity avoidance of all features and linking consonants, therefore allowing us to tease apart their relative contributions.

Accordingly, besides probing the locality effect, this study also aims to better understand the inventory of features that may be involved in triggering the identity avoidance effect in Turkish emphatic reduplication, and if certain features have a greater effect in this phenomenon in relation to other features. In so doing, we also aim to elucidate the question of how the similarity between the consonants in the base and the LC should be specified, for instance, at the level of the total identity of the consonants, or at the level of the individual features? As will be discussed later in Section 2, this study adopts the methodological approach by Graff and Jaeger (2009) given its ability to help us rigorously examine the nature of both locality and feature specificity, and establish whether locality-sensitive feature-specific identity avoidance constraints are part of the grammar.

To address the questions that revolve around locality and feature-specificity, in this study we conducted an acceptability rating judgement experiment (which has rarely been used for the study of the partial reduplication phenomenon) as opposed to researcher's intuitions or experimentally obtained forced-choice task responses. Among other things, our findings provide support for the view that speakers' grammars have active identity avoidance constraints that operate on specific features (e.g., Gallagher and Coon (2009)) and the strength of their effects is a function of the distance between similar consonants (e.g., Pierrehumbert (1993); Zymet (2014, 2018)). These are in line with previous findings on Arabic (Frisch et al., 2004), Dutch, Aymara, and Javanese (Graff & Jaeger, 2009) which highlighted a gradient feature-similarity-based restriction that is also subject to locality (see also Suzuki (1998)).

In terms of the nature of feature specificity, our results reveal that in the Turkish emphatic reduplication process, the similarity between the consonants in the base and the LCs operates at the segmental level (total identity) as well as the level of individual phonological features (partial identity). Furthermore, we found that not all features participate equally by allowing individual features to be free parameters. Our study also confirms the importance of only some of the features employed in the previous studies, such as [strident], [labial] and [nasal], but not others, such as [coronal], [sonorant], or [continuant]. The important features like [strident] and [labial] and the unimportant features such as [coronal] and [sonorant] are more in line with the cross-linguistic tendencies (see e.g., Bye 2011; Pierrehumbert 1993).

In terms of the effect of locality, the study demonstrates that not only the first two consonants, but all the consonants in the base form contribute to the identity avoidance effect (Zymet 2014), with the strength of the effect decreasing further into the base. Another locality-related factor that studies on assimilation (and dissimilation) have noted is the role of *syllable position* (Rose and Walker 2004) or *syllable-role specific correspondence* (Bennett 2012). This approach states that matching syllable roles might contribute to segments' (dis)similarity. Given this aspect has not been investigated in the context of Turkish emphatic reduplication, our study also examines whether syllable position is a significant factor. Strikingly, it turns out the identity avoidance effect in Turkish is also sensitive to the syllable position. In this regard, it lends support to the view that syllable position might play a role in contributing to segments' (dis)similarity (Bennett 2012; Rose and Walker 2004; Suzuki 1998). In particular, the effect would decay with the distance from the linking consonant and would be enhanced if the consonant in the base matches the constituency of the linking consonant (coda vs onset). For instance, in the case of Turkish, the linking consonant is always a coda[8]. When a base form has the word shape C_1VC_2V, a distance-based

analysis (melodic overwriting) of fixed segmentism (see e.g., Kelepir 1999, 2000; Yu 1999).

[8]Note again that the linking consonant is always a coda in the C-initial forms.

decay effect was found, as hypothesised, with C_2 being less influential than C_1. However, the word shape C_1VC_2 exhibits the opposite pattern with C_1 being less influential than C_2. This surprising effect can be attributed to that the LC and the C_2 in C_1VC_2 are both codas (see Section 3.2.2 for the complete result). Accordingly, the Turkish phenomenon is particularly insightful in this regard too: it showcases the interplay of distance-based decay effect and syllable position effect, which in most studies are treated separately (due to the data involved). Moreover, unlike other examples that illustrate the syllable position effect, the presence of this effect in Turkish is not immediately clear both due to the just-mentioned interplay with distance-based decay effect and also because the dissimilation is not categorical (cf. Bennett 2013), but much more gradient.

This study also replicates some of the previous Turkish-specific findings. For example, the preference hierarchy regarding the choice of the LC ([p] > [s] > [m] > [r]) still broadly holds. Methodologically we demonstrate that the precise nature of the identity avoidance effect can be revealed using hierarchical regression and statistical model comparisons (Graff & Jaeger, 2009; Zymet, 2019).

The rest of the paper is organised as follows: Section 2 outlines the details of our study. We first introduce the logic of this study before we present the materials (Section 2.1), followed by the methodological details of the experiment (Section 2.2). Section 2.3 introduces the variables of interest. Section 2.4 outlines the modelling procedure. In this section, we will introduce the model specification and evaluation as well as the modelling details of Study I on feature specificity (Section 2.4.2) and Study II on locality (Section 2.4.3). Section 3 presents our results for the two studies. Section 4 contextualises our findings of the Turkish case into the broader literature, focusing on feature specificity and locality, particularly the interaction between distance-based decay and the syllable position. This section also discusses issues such as factors beyond identity avoidance and representation of speakers' knowledge. Section 5 summarises and concludes the paper.

2. The present study

The present work consists of two studies. Study I addresses the level of similarity on which the identity avoidance effect operates. Study II addresses whether the proximity between the consonants and the linking consonant plays a role in the identity avoidance effect.

To examine these research questions, we conducted a large-scale rating study of 162 real base forms of Turkish sampled from previous studies. Table 1 summarises most of the previous studies along the lines of various criteria that will be referred to in the current study: (i) whether the study relies on the researcher's intuition or an experiment, (ii) what type of experiment was conducted and the number of participants, and (iii) whether the items used are real words or nonce-words.

Table 1. A summary of the data examined in 13 previous studies. FC stands for *forced choice*, and OSR stands for *open-set response*.

Sources	Intuition	Experiment	Type of Exp.	# of Participants	# of Items and Types
Hatiboğlu (1973)	Yes	No	-	-	Real (142)
Demircan (1987)	Yes	Yes	FC	100	Real (110), Nonce (20)
Dobrovolsky (1987)	Yes	No	-	-	Real (9)
Taneri (1990)	-	Yes	FC	32	Real (300)
Wedel (1999)	Yes	Yes	FC	3-8	Real (125 + 80)
Yu (1998)	No	Yes	Rating	4	Real (101), Nonce (56)
Yu (1999)	Yes	No	-	-	Real (152)
Kelepir (2000)	Yes	No	-	-	Real (89)
Sofu (2005)	-	Yes	FC	25 adults, 89 children	Nonce (38)
Sofu and Altan (2008)	-	Yes + Corpus	FC	80	Real (132)
Kaufman (2014)	-	Yes	1-FC, 2-Rating	1-16, 2-50	1-Nonce (44), 2-Real (45)
Demir (2018)	-	Yes	OSR	125	Real (10), Nonce (34)
Köylü (2020)	-	Yes	FC	14	Nonce (48)

The decision of using a rating task was motivated by our desire to fill a methodological gap that was found in previous studies. As summarised in Table 1, in many studies, the judgements were often based on the researcher's intuitions only and whenever there is an experimental component, the task is almost exclusively a forced-choice task. The use of acceptability rating might have an advantage over a forced-choice task in that the forced-choice task might be masking potential variability within participants that we noted in (5).

Note also that previous studies on Turkish either completely ignored the variable nature of the LC, or incorporated into their analyses only the most dominant LC (even when they have empirical basis for the variability). For instance, Wedel (2000) reported that most participants responded with only one LC for each base form, even though the participants as a group chose multiple LCs. In a post experiment interview, the participants reported that the other forms that they did not choose were also possible; however, they simply selected the first one that they had in mind. A rating task would therefore be able to better examine base forms that have a high level of variability, allowing for multiple LCs to apply to the same base form by a given participant.

Our study reports on a large-scale rating experiment with 162 real base forms with at least 40 participants rated each form. It is a methodological improvement over existing studies that made use of a rating task, since they either tested a small number of participants such as Yu (1998) with four participants, or tested a small number of items such as Kaufman (2014) with 10 real base forms.

2.1. Materials

The experimental items were 162 real base forms taken from previous studies to enhance the comparability and the replicability of this work. Most of the items were taken from the classic study by Hatiboğlu (1973) because many of the later studies also examined a subset of these items, and the rest were sampled from the other studies.[9] The 162 items were then evenly divided into five lists (three lists have 33 items; one list has 32 items; and one list with 31 items) with each list containing roughly the same distribution of dominant linking consonants as well as variable items.[10]

[9]In addition to the items from Hatiboğlu (1973), we included a few more items from the list in Stachowski (2014), which compiled the list of mostly overlapping items from Hatiboğlu (1973), Demircan (1987), Müller (2003), and others. Out of the 178 items in Stachowski's (2014) list, we left out 16 items since some of them were nouns (e.g., *buz* 'ice', or *çevre* 'environment'), and others were items not available in Turkey Turkish, but other Turkic varieties such as Azeri Turkish (e.g., *deyirmi* 'circle').

[10]The expected dominant linking consonant of each item was based on previous studies as well as the linguistic intuition of Faruk Akkuş, the co-author who is a native Turkish speaker. The expected distribution of the items

2.2. Methods

Each participant was asked to perform both a rating task and a forced-choice task (not reported here)[11]. For each base form, all four of its reduplicated forms (each with a different linking consonant (LC)) were shown on the same screen orthographically. Each participant was randomly assigned to one of the five lists. The order of the two tasks and the order of the four reduplicated forms were also randomised for each participant. Each reduplicated form was rated on a scale of naturalness: DOĞAL DEĞİL 'not natural' [1 to 7] DOĞAL 'natural'. The experiment was programmed using Experigen (Becker & Levine, 2013), hosted at `http://db.phonologist.org/`.

The experiment was advertised via social media. Participants were invited to take part in the experiment voluntarily and given informed consent. The inclusion criteria of our target population were native Turkish speakers, born in Turkey without language-related disorders. A total of 283 participants completed the experiment. 207 participants who met our inclusion criteria were included in the analyses. All items were rated by at least 40 participants.[12]

We evaluated the results of the rating study by adapting Graff and Jaeger's (2009) methodological approach. Graff and Jaeger (2009) examined the feature specificity and locality of the identity avoidance effect in the lexical organisation of Aymara, Dutch and Javanese. Methodologically they made use of a regression approach to allow for individual identity factors as well as nuisance factors to act as free parameters. They compared the different types of identity factors (total identity and featural identity between two segments) and whether these factors are strictly local by using a model comparison approach. Their key findings were that i) both the total identity and the featural identity of two segments influence the formation of lexical roots, ii) the identity avoidance effect operates over individual features which have their individual weights, and iii) the identity avoidance effect operate both locally and non-locally. Our study aims to ask the same research questions concerning the identity avoidance effect. Using this approach we were able to quantitatively examine the effect of feature specificity and locality rather than heuristically as was done in many previous studies. Unlike the previous studies which modelled only the dominant linking consonant of each base form, we modelled trial-level responses from each participant using linear mixed-effects regression.

While the individual features included as predictors are also the same set of features that suffers from the same feature redundancy issue discussed in Section 1.3, the statistical regression approach enables researchers to deal with this issue by disentangling the direct effects of specific variables. The coefficient of each variable in a multiple regression model represents the variable's influence while statistically controlling for the influence of other variables (Winter, 2019). Thus this study makes use of a novel approach of using statistical reasoning to determine which features are relevant, rather than preselecting a particular set of features using more traditional phonological means.

The 162 items were chosen to ensure that words with more than two consonants are well represented (see Table 3 for the breakdown of the selected words by the number of consonants). This was motivated by how previous studies were restricted in their ability to properly examine the potential contribution from consonants further away from the reduplicant due to limited number of stimuli with more than two/three consonants limit (see Section 1.1 for the full discussion).

consists of 55% P-items, 9% M-items, 17% S-items, 15% R-items and 24% variable items. This information was not used in our statistical analyses and served merely for the purpose of the experiment design.

[11] A simple item-level correlation analysis suggests the judgements from the forced-choice task are similar to those obtained from the rating task (R^2: 0.85).

[12] Their mean age was 27.44 years, ranged between 18 and 63. 146 were women and 61 were men.

These items are part of the existing vocabulary of the Turkish lexicon. Therefore, they are likely to have an uneven representation of consonants and their features. Due to this uneven representation in the items, it is possible that, for example, C_1 and its features might not have the same effect across the stems with different numbers of consonants. The nature of the stimuli therefore imposes a limit as to how strong an effect of a given feature can exert; for example, [strident] can be shown to be significant across items but [labial] may not be because its representation happens to be low. Not separating the items by the number of consonants could mask these potential differences (as we will see in the result section, there are indeed notable differences; see Section 4 for a discussion).[13] We therefore chose to model our items separately for stems with different numbers of consonants (see Sections 2.4.2 and 2.4.3 for details). Furthermore, this analytical approach is necessitated by the regression modelling approach by Graff and Jaeger (2009) that we are adopting. For instance, if we were to combine both items with two and three consonants in a single regression model, then the predictors concerning the third consonants would not be able to be specified for the items with only two consonants. Relatedly, given the nature of the dataset and the modelling procedure, it is worth noting what can be inferred from the results. First, we can infer the overall importance of the features in Turkish emphatic reduplication by examining how prevalent they are across models in terms of their role in capturing the data. Second, we can also infer the nature of feature specificity by comparing between models (using model comparisons) with different formulations of identity avoidance predictors. Third, we can also infer the nature of locality by comparing models (again with model comparison) with and without the features associated with a specific consonant. Please see Section 4.1 for a detailed discussion.

2.3. Variables

This section describes the fixed effect variables and the random effect variables we included in our analyses. Variables that we excluded can be found in Appendix B.

2.3.1. Fixed effect variables

Total identity: Each consonant in the base is encoded for whether it is identical to the LC with non-identical being the reference level. For instance, a base with two consonants have two binary variables. This variable is a version of the previously proposed constraint such as full reduplication or no-repeat which checks the total identity of C_2 and the LC; however, it generalises across all consonants and not only C_2.
Partial identity: Each consonant in the base is encoded for whether the feature value of each of its phonological features is identical to that of the LC with non-identical being the reference level. It is important to note that only positive featural values are compared. Following the phonological system outlined in Erguvanlı Taylan (2015), eight binary consonantal features were used: sonorant, voice, continuant, strident, anterior, coronal, labial, and nasal (see Appendix A for the feature chart).[14] The features, high, back and lateral, were excluded because all of the four linking consonants have a negative value for these features. The partial identity was modelled using two approaches. The first approach is to

[13] The use of existing words raises the issue of lexicalization, which is addressed in Section 4.4 and argued to not be the case. Ultimately, even if this was a matter of lexicalization, our study would provide a more delicate measure of how much lexicalization affects people's sensitivity in their choice due to rating task we employ.

[14] Following Erguvanlı Taylan (2015), the so-called 'soft-g' ğ was treated as a voiced velar fricative /ɣ/ in the calculation of partial identity. Since only nine out of 162 items contain the 'soft-g', we anticipate that a different methodological decision should have a negligible effect on our findings. We encourage readers to experiment with different feature systems using our data and scripts on osf.io.

allow each of the eight matched features to contribute differently by using them as individual variables. This approach would create eight binary variables for each of the consonants in the base form. We refer to these variables as *individual feature identity*. The second approach is to sum up the number of matched features, thus assuming that all features have the same weight. This would yield one continuous variable for each of the consonants in the base form. We refer to this variable as *sum featural identity*.

Transitional phonotactic probability: Demircan (1987) proposed speakers might avoid selecting a linking consonant that would lose or change the distinctive features of C_1 due to the principle of least effort. For instance, consonantal sequences across syllables such as [p.b] might undergo devoicing which makes the base less intelligible, therefore [p] is unlikely to be selected in that context. Similarly, Wedel (2000) observed that plosives are generally dispreferred in Turkish phonotactics if followed by a homorganic consonant since they are articulatorily or perceptually marked. The authors confirmed this observation by conducting a corpus search using an online Turkish lexicon (TELL) (Inkelas, Küntay, Sprouse, & Orgun, 2000).

Following Wedel (2000), lexical statistics were used to quantify the degree of junctural markedness. The assumption is that speakers are unlikely to produce consonant clusters that are articulatorily difficult or perceptually less distinctive. The token frequencies of all intervocalic heterosyllabic two-consonant clusters beginning with one of the four LCs ([Vp.CV],[Vm.CV],[Vs.CV] and [Vr.CV]) were extracted from a large subtitle-based corpus of Turkish. The written corpus was compiled using over 40,000 subtitle texts of Turkish. The corpus contains approximately 200 million word tokens and over 200,000 word types. The use of a subtitle corpus was motivated by the fact that lexical frequencies derived from subtitle texts have consistently shown to outperform those from other genres in capturing behavioural responses in psycholinguistic tasks across languages (Brysbaert & New, 2009; Keuleers, Brysbaert, & New, 2010; Tang, 2012; Tang & de Chene, 2014). The expectation is that the higher the transitional phonotactic probability (estimated using token frequency) of a juncture sequence, the higher the acceptability rating (Albright, 2007; Bailey & Hahn, 2001; Goldrick, 2011).

2.3.2. Coding illustration of identity variables

Table 2 illustrates how the identity variables are coded for C_1 [s] of the stimulus *sarı* 'yellow'. The total identity of C_1 matches with the LC [s] and not with the LCs [p,m,ɾ], therefore the total identity variable with LC [s] is coded as 1 and the others are coded as 0s. Concerning the individual featural identity variables, C_1 and the LC [ɾ] both have a positive feature value for the features [continuant], [anterior], [cororal], therefore these features are coded as 1s, while the other features ([sonorant], [voice], [strident], [labial] and [nasal]) are coded as 0s. The sum featural identity variable for the LC [ɾ] is the sum of the number of matched features which is 3 ([continuant], [anterior] and [cororal]).

Table 2. Illustration of identity variables of C_1 for the stimulus *sarı* 'yellow' with each of the four linking consonants. Abbreviations: son: sonorant, cont: continuant, strid: strident, ant: anterior, cor: coronal, lab: labial, lat: lateral, and nas: nasal. The individual feature identity variables are son, voice, cont, strid, ant, cor, lab, and nas. sum stands for the sum featural identity; total stands for the total identity variable.

sarı 'yellow'			Matching C_1									
C_1	C_2	LC	son	voice	cont	strid	ant	cor	lab	nas	sum	total
s	r	p	0	0	0	0	1	0	0	0	1	0
s	r	m	0	0	0	0	1	0	0	0	1	0
s	r	s	0	0	1	1	1	1	0	0	4	1
s	r	r	0	0	1	0	1	1	0	0	3	0

2.3.3. Random effect variables

As is typical of psycholinguistic research, participant and base form were included as random effects to allow for idiosyncrasies of individual participants and items.

Prior studies have established that only the consonants of the base trigger the identity avoidance effect in Turkish. Therefore, we analysed any base words that have the same number of consonants together. However, to recognise that these base words do in fact have different shapes (different number of syllables, and closed vs open syllables in different positions), we encoded the word shape of the base forms (such as CVC and CVCV, etc. as shown in Table 3) as a random effect to capture its potential effect. These models thus allow us to focus on the factors of interest, i.e., feature mismatch.[15] The linking consonant of the reduplicated forms was also included as a random effect to capture the general preference for specific linking consonants.

2.4. Modelling procedure

The 162 items consists of 27 vowel-initial items and 135 consonant-initial items. Vowel-initial items were not analysed since they have an overwhelming preference for the LC [p] (Demircan 1987; Kelepir 1999; Sofu 2005; Sofu and Altan 2008, i.a.). This preference for [p] is supported by the descriptive statistics of the ratings of the 27 items. The mean and median ratings (on a scale of 1 to 7) are 6.402 and 7 for [p], 1.502 and 1 for [m], 1.533 and 1 for [s], and 1.197 and 1 for [r]. The 135 consonant-initial items were divided up into four groups based on the number of consonants they contain in the base and their word shapes as shown in Table 3. The five-consonant items were filtered out because they were only six of them and they might not provide enough statistical power for the analyses.

2.4.1. Model specification and evaluation

Linear mixed-effects regression models were fit to the rating responses conducted using the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2013). Following standard practice in regression modelling, the continuous variables were z-score normalised (e.g., Baayen, 2008, Sec. 2.2). Z-score normalization allows us to compare the relative strength of our continuous predictors directly. As per standard practice with token

[15]Note that the inclusion of word shape as a random effect was only for the models for which word shape was not the variable of interest. In Study II, when examining the effect of word shape, we constructed a separate regression model for each word shape, therefore word shape was not included as a random effect.

Table 3. A summary of the stimuli categorised by the number of consonants in the base and their word shapes. The number in parentheses indicates the number of items for each word shape.

C_1C_2	$C_1C_2C_3$	$C_1C_2C_3C_4$	$C_1C_2C_3C_4C_5$
C_1VC_2 (23)	$C_1VC_2VC_3$ (37)	$C_1VC_2C_3VC_4$ (20)	$C_1VC_2VC_3C_4VC_5$ (5)
C_1VC_2V (19)	$C_1VC_2C_3V$ (14)	$C_1VC_2VC_3VC_4$ (8)	$C_1VC_2VC_3VC_4C_5$ (1)
	$C_1VC_2C_3$ (4)	$C_1VC_2VC_3C_4V$ (1)	
	$C_1VC_2VC_3V$ (2)	$C_1VC_2C_3VC_4V$ (1)	

frequency, the transitional phonotactic probability was log-transformed (base 10) before z-score normalization. Our categorical predictors was sum-coded (Wissmann, Toutenburg, & Shalabh, 2007) with non-identical as the base level.

The statistical significance of the individual predictors in all the models was evaluated by bootstrapping. Bootstrapping was carried out using the *bootmer* function in the *lme4* library. 1,000 bootstrap simulations were performed for each model. Bootstrapped p-values and confidence intervals at 95% were computed for each predictor in each model. We follow the conventional alpha-level of 0.05 for significance. Model comparisons were performed using Akaike information criterion (AIC).[16] All models underwent the process of model criticism. For each model, the residuals were extracted and data points that were 2.5 standard deviations above or below the mean residual value were excluded. No more than 1% of the data points were excluded in any of the models.

To evaluate potential collinearity issues, we computed the Variance Inflation Factor (VIF) of the predictor variables in each of the models. The variables in all but four of the models have VIF < 5. These four models have in total 14 variables that have VIF > 5 but < 10, and 2 variables have VIF slightly above 10. These values are mostly below the typical critical values of 5 or 10 (Chatterjee & Hadi, 2015; Tomaschek, Hendrix, & Baayen, 2018) which indicates no serious issues of collinearity.[17]

In this study, we make use of zero-order correlations to address the issues of interpretability due to potential collinearity, even if the potential of serious collinearity is low as suggested by the VIF analyses.[18] The correlation amongst the identity predictors could still cause their effects to be counterintuitive and hard to interpret. Collinearity between two predictors can cause the reduction or sign reversal in one of the model estimate. For instance, some of the identity predictors might behave in the opposite direction of identity avoidance with their regression coefficients being positive. One diagnostic of a suppressor effect is whether the model estimate is in the same or opposite direction as the correlation between the dependent and independent variable. Model estimates in the opposite direc-

[16]Model comparisons were also performed using Bayesian information criterion (BIC). The penalty term for the number of parameters is larger in BIC than in AIC. BIC is useful when the number of parameters between the two models being compared is particular different. All the results were the same using AIC or BIC, therefore only the results using AIC were reported.

[17]Two of these four models have only 2 variables with a VIF above 5 but below 6 with the maximum VIF of 5.79 and 5.87. One of the models has five variables with a VIF above 5 but below 10 and one variable with a VIF of 10.2. The remaining model has seven variables with a VIF above 5 but below 10 and one variable with a VIF of 10.5. The variables associated with these higher VIFs are all individual featural identity variables. It is worth noting that recommendations of VIF's cutoff vary depending on the literature, e.g., Hepworth, Gordon, and McCullough (2007) suggests 4 to be the cutoff. It is not clear what a meaningful boundary is for a low versus a high value. No fixed set of guidelines can guarantee the correct analysis of collinear data (Tomaschek et al., 2018). We acknowledge that the best practice of dealing with collinearity has not been established.

[18]We have made our data and analysis scripts available on osf.io and we encourage readers to evaluate the data and our procedures themselves and to examine with different statistical techniques.

tion of the correlation suggest a suppressor effect. Follow-up inspections were performed by examining the pairwise zero-order association between each of individual predictors and the response variable. Pairwise zero-order association is to estimate the effect of individual independent variables has on the dependent variable (see Appendix D for discussion). This was done by fitting multiple mixed-effects regression models with only one independent variable at a time. A regression model with the same random effect structure as the above models was fitted with only each of the predictors for each of the three groups of base forms. A null model with no fixed effect variables was fitted to compare with each of these models with one fixed effect variable. The drop in AIC values was used as a measure of the importance of each variable (AIC_{null} - $AIC_{superset}$). A drop in AIC of more than 2 indicates statistical significance. When interpreting the direction of the identity effects in the full models, these pairwise associations would assist us in identifying cases of a sign reversal due to collinearity.

The distribution of the variables (both the response variable and the predictors) for each of the three item groups with two, three and four consonants in the base form can be found in Appendix C. Means and standard deviations of the by-participant standardised ratings of each item (a base form with one of the four linking consonants) can be found in Appendix H. The complete report of the statistical analyses (regression tables, model comparisons, and figures) can be found on the `osf.io` repository (see Section 5 *Data accessibility statement*).

2.4.2. Study I: Feature specificity

Study I focuses on three groups of items: 42 two-consonant base forms, 57 three-consonant base forms and 30 four-consonant base forms. The analyses were conducted separately for each of the three item groups. Two full models were initially fitted. The two models differ in the type of partial identity variables. One model has individual feature identity variables (eight binary variables per consonant in the base), while the other has sum featural identity variables which are the number of identical features and are computed by summing up the number of identical features, thus assuming that all features have the same weight (one continuous variable per consonant in the base). To assess the level of feature specificity, a series of model comparisons was performed by removing each type of identity variables in bulk. Three more subset models were therefore fitted: a) A model with total identity variables without partial identity variables, b) a model with individual featural identity variables without total identity variables, c) a model with sum featural identity variables without total identity variables. These models were fitted with the predictor variables outlined in Section 2.3.1 as fixed effects and four random intercepts with the variables outlined in Section 2.3.3.

The regression structures of the two full models are shown below. Note that the identity variables have C_i in parentheses and the index i is referring to a specific consonant in the base form. If the base form contains N consonants, then there would be N sets of identity variables.[19]

Model with total identity and partial identity using individual features:

> Rating $\sim$ Total identity (C_i) + Sonorant identity (C_i) + Voice identity (C_i) + Continuant identity (C_i) + Strident identity (C_i) + Coronal identity (C_i) + Labial identity (C_i) + Nasal identity (C_i) + Transitional phonotactic probability + (1 | Participant) + (1 | Base form) + (1 | Word shape) + (1 | Linking consonant)

Model with total identity and partial identity using the sum of the matched features:

[19]See Section 2.2 for the rationale behind the modelling approach.

Rating $\sim$ Total identity (C_i) + Sum featural identity (C_i) + Transitional phonotactic probability + (1 | Participant) + (1 | Base form) + (1 | Word shape) + (1 | Linking consonant)

2.4.3. Study II: Locality: distance-based decay and syllable role

Study II consists of two analyses. The first analysis aims to address the importance of the consonants beyond C_2 (namely C_3 and C_4). The second analysis focuses on examining how the identity avoidance effect would be affected if the consonant in the base matches the constituency of the linking consonant.

In the first analysis, the initial models were the best models found in Study I using model comparisons. Model comparisons were performed by removing identity variables (total and partial) that are associated with each consonant position in bulk. The drop in AIC values were used as a measure of the importance of the consonant (AIC_{subset} - $AIC_{superset}$).

In the second analysis, the best model structure in the first analysis was fitted over base forms with each of the word shapes separately without the random variable (word shape). The five most frequent word shapes were selected since they have a relatively higher number of base forms (at least 14) to enhance the generalisability of our findings. The structures are C_1VC_2, $C_1VC_2VC_3$, $C_1VC_2C_3VC_4$, C_1VC_2V and $C_1VC_2C_3V$. Identity variables (total and partial) that are associated with each consonant position were dropped in bulk and the drop in AIC values were computed. The variable importance values between each of the consonants within each word shape were compared to enable an examination of how syllabification of the consonants plays a role, specifically whether the consonant in the base matches the constituency of the linking consonant.

3. Results

3.1. Study I: Feature Specificity

In the following sections, we present the results from the model comparisons of different levels of feature specificity in Section 3.1.1. Given the model comparisons, the selected best models were then evaluated further individually for each of the three item groups. To enhance the interpretability of each of the predictors in the best models, all pairwise associations between the response and each of the predictors were computed and can be found in Appendix D. Finally we report the detailed model evaluations of each of the predictors in Sections 3.1.2, 3.1.3 and 3.1.3 for the two-consonant, three-consonant, and four-consonant groups respectively. The regression tables of all of the models can be found on the `osf.io` repository (see Section 5 *Data accessibility statement*).

3.1.1. Model comparison

To evaluate the level of specificity of our identity variables, five models were fitted and evaluated for their AIC levels for each of the three item groups (base forms with either two (C_1C_2), three ($C_1C_2C_3$) or four ($C_1C_2C_3C_4$) consonants). These five models include the two full models (total identity and sum featural identity; total identity and individual featural identity) and three subset models without either the total identity variables or the partial identity variables (total identity, sum featural identity, individual featural identity). The AIC values of all models are summarised in Table 15 (See Section F in the Appendix).

The model structure with both the total identity variables and the individual featural identity variables consistently yielded lower AICs (the best model fit) across the three item groups (two-consonant: 28853.63, three-consonant: 39131.32 and four-consonant:

20298.78). This finding suggests that both total identity and partial identity play a role in identity avoidance in Turkish partial reduplication. This supports many of the previous analyses which take into account of both total and partial identity avoidance, for instance, Demircan (1987)'s observations that the linking consonant (LC) cannot be identical with any of the consonants in the base and the features of the LC should not be identical to the C_2 of the base. This finding is also in line with how both total and partial identity influence consonant co-occurrence patterns within lexical roots (Gallagher & Coon, 2009; Graff & Jaeger, 2009).

Given that both total identity and and partial identity are important, their relative importance is also examined. Their relative importance can be evaluated by comparing the drop in AIC values when either of these variable types was dropped from a full model. This was computed separately the two full model structures (total identity and sum featural identity; total identity and individual featural identity). The drop in AIC is summarised in Table 14 (See Section F in the Appendix). The individual featural identity has a bigger drop in AIC than total identity in the two-consonant group (1016.18 vs 87.07), the three consonant-group (1791.52 vs 291.94), and the four-consonant group (1094.96 vs 249.08). Similarly, the sum featural identity has a bigger drop in AIC than total identity in the two-consonant group (461.79 vs 150.88), the three consonant-group (1028.13 vs 376.39), and the four-consonant group (473.5 vs 339.42). This finding suggests that partial featural identity (sum or individual) has a stronger effect on the linking consonant than total identity. This, as far as we know, has not been formally established in previous studies of Turkish partial reduplication.

3.1.2. Model evaluation: Two-consonant group

The fixed and random effects estimates of the two-consonant model are summarised in Table 4 and Table 11 (see Appendix E), respectively. First of all, the transitional phonotactic probability was not statistically significant ($p = 0.514$). We turn now to the identity variables, starting with those in C_1. All but the continuant identity variable ($p = 0.750$) and the anterior identity variable ($p = 0.082$) were statistically significant. An identity avoidance effect was found in five of the seven significant identity variables with a negative coefficient – total identity ($\beta = -1.6748$), sonorant identity ($\beta = -1.4779$), strident identity ($\beta = -1.7381$), coronal identity ($\beta = -0.6159$), and labial identity ($\beta = -1.2808$). The remaining two signficant variables, voice identity ($\beta = 1.3000$) and nasal identity ($\beta = 2.2870$), have a positive coefficient, suggesting the opposite effect of identity avoidance. To clarify these two variables, we turn to the pairwise association analysis in Table 10. The voice identity variable has a genuine identity *preference* effect ($\beta = 1.3000$) since the zero-order association is also positive (Table 10, $\beta = 0.4045$). The nasal identity variable shows a suppressor effect since the coefficient is positive ($\beta = 2.2870$) even though the zero-order association is negative (Table 10, $\beta = -2.2434$).

We now focus on the identity variables in C_2. All but three variables were statistically significant. The insignificant variables are anterior identity ($p = 0.262$), coronal identity ($p = 0.336$) and nasal identity ($p = 0.504$). An identity avoidance effect was found in five of the six significant identity variables with a negative coefficient – sonorant identity ($\beta = -0.7207$), voice identity ($\beta = -1.7778$), continuant identity ($\beta = -0.8210$), strident identity ($\beta = -1.2272$) and labial identity ($\beta = -3.6485$). The total identity variable shows a suppressor effect since the coefficient is positive ($\beta = 0.7255$) even though the zero-order association is negative (Table 10, $\beta = -0.8473$).

Table 4. Fixed effects summary for Study I (two-consonant base forms). β: coefficient; SE: standard error; t: t-value; $\text{CI}_{Lower\ 95\%}$ and $\text{CI}_{Upper\ 95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations. Significant variables ($p \leqslant 0.05$) are in bold.

		β	SE	t	$\text{CI}_{Lower95\%}$	$\text{CI}_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	-0.7366	0.5939	-1.2402	-1.9217	0.4272	0.252
C_1	**Total identity**	-1.6748	0.2209	-7.5826	-2.1176	-1.2124	$<.001^{***}$
	Sonorant identity	-1.4779	0.2098	-7.0449	-1.8832	-1.0523	$<.001^{***}$
	Voice identity	1.3000	0.1485	8.7545	1.0002	1.5936	$<.001^{***}$
	Continuant identity	0.0378	0.1297	0.2915	-0.2043	0.2868	0.750
	Strident identity	-1.7381	0.2278	-7.6310	-2.2085	-1.2861	$<.001^{***}$
	Anterior identity	0.5291	0.2961	1.7870	-0.0577	1.1192	$0.082^{\cdot}$
	Coronal identity	-0.6159	0.1130	-5.4507	-0.8271	-0.3867	$<.001^{***}$
	Labial identity	-1.2808	0.1327	-9.6542	-1.5404	-1.0173	$<.001^{***}$
	Nasal identity	2.2870	0.3653	6.2600	1.5439	3.0079	$<.001^{***}$
C_2	**Total identity**	0.7255	0.1217	5.9596	0.4904	0.9635	$<.001^{***}$
	Sonorant identity	-0.7207	0.1258	-5.7281	-0.9809	-0.4719	$<.001^{***}$
	Voice identity	-1.7778	0.1211	-14.6801	-2.0168	-1.5447	$<.001^{***}$
	Continuant identity	-0.8210	0.0996	-8.2377	-1.0144	-0.6254	$<.001^{***}$
	Strident identity	-1.2272	0.1399	-8.7738	-1.4799	-0.9418	$<.001^{***}$
	Anterior identity	0.3377	0.2854	1.1833	-0.2387	0.9052	0.262
	Coronal identity	-0.1035	0.1054	-0.9810	-0.3146	0.1180	0.336
	Labial identity	-3.6485	0.2078	-17.5598	-4.0599	-3.2392	$<.001^{***}$
	Nasal identity	0.1762	0.2763	0.6375	-0.3666	0.7075	0.504
	Transitional phonotactic probability	-0.0408	0.0594	-0.6859	-0.1596	0.0773	0.514
Number of observations: 6,883; number of participants: 207; number of base forms: 42; number of word shapes: 2; number of linking consonants: 4							
Level of significance: $\cdot$ ($p \leqslant 0.1$), * ($p \leqslant 0.05$), ** ($p \leqslant 0.01$), *** ($p \leqslant 0.001$).							

3.1.3. Model evaluation: Three-consonant group

The fixed and random effects estimates of the three-consonant model are summarised in Table 5 and Table 12 (see Appendix E), respectively. First of all, the transitional phonotactic probability was statistically significant in the positive direction (β = 0.1821, p = 0.514), suggesting a preference for an articulatorily or perceptually unmarked heterosyllabic cluster. We turn now to the identity variables, starting with those in C_1. An identity avoidance effect was found in all of the five significant identity variables with a negative coefficient – total identity (β = -1.8850), voice identity (β = -0.4231), continuant identity (β = -0.5254), strident identity (β = -1.6904), and labial identity (β = -2.2846). The remaining three variables were not significant and they are sonorant identity (p = 0.312), anterior identity (p = 0.350) and coronal identity (p = 0.092). Nasal identity was excluded in this model because the base forms have no nasals in C_1.

We now focus on the identity variables in C_2. All but one variable were statistically significant. The insignificant variable is sonorant identity (p = 0.118). An identity avoidance effect was found in six of the eight significant identity variables with a negative coefficient – voice identity (β = -2.0183), continuant identity (β = -0.6266), strident identity (β = -1.3080), cororal identity (β = -1.1598), labial identity (β = -2.2197) and nasal identity (β = -1.0419). The total identity variable shows a suppressor effect since the coefficient is positive (β = 1.2713) even though the zero-order association is negative (Table 10, β =

-0.9739). The positive coefficient (β = 0.4863) of the anterior identity variable is unlikely to be genuine because its level of significance is weak with a $p - value$ of 0.024 and while the zero-order association is also positive (β = 0.0160), it was insignificant with a small effect size (a featural match increases the rating by only 0.016 on a scale from 1 to 7).

We now turn to the identity variables in C_3. All but the total identity variable (p = 0.1) and the anterior identity variable (p = 0.1) were statistically significant. An identity avoidance effect was found in four of the seven significant identity variables with a negative coefficient – sonorant identity (β = -1.3457), strident identity (β = -1.2153), coronal identity (β = -0.8877) and nasal identity (β = -1.6972). The voice identity variable and the labial identity variable both show a suppressor effect since the coefficients are positive (voice: β = 0.9853; labial: β = 0.8738) even though the zero-order associations are negative (Table 10, voice: β = -0.5918 and labial: β = -1.5442). The continuant identity variable potentially has a genuine identity preference effect (β = 0.2059) since the zero-order association is also positive but only near-significant (Table 10, β = 0.1514, ΔAIC = 1.63).

Table 5. Fixed effects summary for Study I (three-consonant base forms). β: coefficient; SE: standard error; t: t-value; $CI_{Lower\ 95\%}$ and $CI_{Upper\ 95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations. Significant variables ($p \leqslant 0.05$) are in bold.

		β	SE	t	$CI_{Lower95\%}$	$CI_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	-2.6151	0.4537	-5.7641	-3.5669	-1.7126	<.001***
C_1	**Total identity**	-1.8850	0.1361	-13.8534	-2.1466	-1.6117	<.001***
	Sonorant identity	0.1596	0.1492	1.0697	-0.1345	0.4529	0.312
	Voice identity	-0.4231	0.1118	-3.7850	-0.6471	-0.2043	<.001***
	Continuant identity	-0.5254	0.0933	-5.6289	-0.7022	-0.3440	<.001***
	Strident identity	-1.6904	0.1259	-13.4288	-1.9466	-1.4399	<.001***
	Anterior identity	-0.2035	0.2223	-0.9155	-0.6491	0.2353	0.350
	Coronal identity	0.1700	0.0971	1.7504	-0.0237	0.3658	0.092·
	Labial identity	-2.2846	0.1264	-18.0688	-2.5308	-2.0290	<.001***
	Nasal identity	–	–	–	–	–	–
C_2	**Total identity**	1.2713	0.1158	10.9754	1.0434	1.5075	<.001***
	Sonorant identity	0.1525	0.0980	1.5552	-0.03312	0.3422	0.118
	Voice identity	-2.0183	0.1149	-17.5622	-2.2374	-1.8008	<.001***
	Continuant identity	-0.6266	0.0951	-6.5911	-0.8161	-0.4352	<.001***
	Strident identity	-1.3080	0.1154	-11.3340	-1.5395	-1.0837	<.001***
	Anterior identity	0.4863	0.2242	2.1691	0.0412	0.9340	0.024*
	Coronal identity	-1.1598	0.0976	-11.8878	-1.3522	-0.9646	<.001***
	Labial identity	-2.2197	0.1142	-19.4466	-2.4463	-1.9995	<.001***
	Nasal identity	-1.0419	0.1487	-7.0058	-1.3387	-0.7497	<.001***
C_3	Total identity	-0.2254	0.1453	-1.5512	-0.5064	0.0480	0.1·
	Sonorant identity	-1.3457	0.1424	-9.4516	-1.6230	-1.0667	<.001***
	Voice identity	0.9853	0.1337	7.3689	0.7202	1.2473	<.001***
	Continuant identity	0.2059	0.0807	2.5491	0.0438	0.3747	0.016*
	Strident identity	-1.2153	0.1232	-9.8688	-1.4585	-0.9716	<.001***
	Anterior identity	0.3722	0.2261	1.6464	-0.07387	0.8059	0.1·
	Coronal identity	-0.8877	0.0899	-9.8803	-1.0608	-0.7172	<.001***
	Labial identity	0.8738	0.1722	5.0740	0.5124	1.2165	<.001***
	Nasal identity	-1.6972	0.1410	-12.0394	-1.9702	-1.4165	<.001***
	Transitional phonotactic probability	0.1821	0.0420	4.3398	0.1017	0.2628	<.001***

Number of observations: 9,374; number of participants: 207; number of base forms: 57; number of word shapes: 4; number of linking consonants: 4
Level of significance: · ($p \leqslant 0.1$), * ($p \leqslant 0.05$), ** ($p \leqslant 0.01$), *** ($p \leqslant 0.001$).

3.1.4. Model evaluation: Four-consonant group

The fixed and random effects estimates of the four-consonant model are summarised in Table 6 and Table 13 (see Appendix E), respectively. First of all, the transitional phonotactic probability was statistically significant in the expected positive direction ($\beta = 0.4553$). We turn now to the identity variables, starting with those in C_1. Only three identity variables were significant and they are all in the negative direction, suggesting an identity avoidance effect – total identity ($\beta = -3.1510$), strident identity ($\beta = -0.8103$) and labial identity ($\beta = -3.0314$). Nasal identity was excluded in this model because the base forms have no nasals in C_1.

We now focus on the identity variables in C_2. All but one variable were statistically significant. The insignificant variable is nasal identity ($p = 0.73$). An identity avoidance effect was found in four of the eight significant identity variables with a negative coefficient

– voice identity (β = -2.7016), continuant identity (β = -1.5544), coronal identity (β = -0.8464) and labial identity (β = -2.1668). The total identity variable and the sonorant identity variable both show a suppressor effect since the coefficients are positive (total: β = 1.1706; sonorant: β = 0.4956) even though the zero-order associations are negative (Table 10, total: β = -0.2840 and sonorant: β = -0.5162). The strident identity variable also shows a suppressor effect since the coefficient is negative (β = -0.8427) even though the zero-order association is positive (Table 10, β = 0.4654). The anterior identity variable has a genuine identity preference effect (β = 1.3207) since the zero-order association is also positive (Table 10, β = 0.4449).

We now turn to the identity variables in C_3. All but the anterior identity (p = 0.618) were statistically significant. An identity avoidance effect was found in four of the eight significant identity variables with a negative coefficient – total identity (β = -0.7889), voice identity (β = -0.6932), labial identity (β = -0.7807) and nasal identity (β = -1.1454). The strident identity variable shows a suppressor effect since the coefficient is negative (β = -2.3943) even though the zero-order association is positive (Table 10, β = 0.9814). Two of the significant identity variables (continuant and coronal identity variables) show a genuine identity preference effect since the coefficients are positive (continuant: β = 0.8278; coronal: β = 0.4544) and the zero-order associations are also positive (Table 10, continuant: β = 1.2100 and coronal: β = 0.6786). The positive coefficient (β = 1.548) of the sonorant identity variable is unlikely to be genuine because while the zero-order association is also positive (β = 0.0436), it was insignificant with a small effect size (a featural match increases the rating by only 0.0436 on a scale from 1 to 7).

Finally we turn to the identity variables in C_4. Four of the variables were not statistically significant – continuant identity (p = 0.616), anterior identity (p = 0.806), coronal identity (p = 0.184) and labial identity (p = 0.296). Five of the variables were statistically significant. An identity avoidance effect was found in four of the five significant identity variables with a negative coefficient – total identity (β = -0.9657), voice identity (β = -0.6780), strident identity (β = -3.4709) and nasal identity (β = -1.7796). The sonorant identity variable has a genuine identity preference effect (β = 1.9742) since the zero-order association is also positive (Table 10, β = 0.2887).

Table 6. Fixed effects summary for Study I (four-consonant base forms). β: coefficient; SE: standard error; t: t-value; $CI_{Lower\ 95\%}$ and $CI_{Upper\ 95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations. Significant variables ($p \leqslant 0.05$) are in bold.

		β	SE	t	$CI_{Lower95\%}$	$CI_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	-5.4181	0.9340	-5.8006	-7.2647	-3.5117	$<.001^{***}$
C_1	**Total identity**	-3.1510	0.2061	-15.2876	-3.5550	-2.7555	$<.001^{***}$
	Sonorant identity	-0.4286	0.2290	-1.8715	-0.8854	0.0268	$0.07^{\cdot}$
	Voice identity	0.4343	0.2426	1.7906	-0.0347	0.9120	$0.074^{\cdot}$
	Continuant identity	0.0156	0.1672	0.0930	-0.3133	0.3449	0.942
	Strident identity	-0.8103	0.2130	-3.8041	-1.2120	-0.3986	$<.001^{***}$
	Anterior identity	0.3652	0.4687	0.7791	-0.5533	1.2584	0.420
	Coronal identity	-0.0721	0.1758	-0.4098	-0.4376	0.2848	0.714
	Labial identity	-3.0314	0.2253	-13.4560	-3.4557	-2.5791	$<.001^{***}$
	Nasal identity	–	–	–	–	–	–
C_2	**Total identity**	1.1706	0.1662	7.0455	0.8619	1.4836	$<.001^{***}$
	Sonorant identity	0.4956	0.1743	2.8434	0.1621	0.8409	0.006^{**}
	Voice identity	-2.7016	0.2063	-13.0954	-3.1069	-2.2986	$<.001^{***}$
	Continuant identity	-1.5544	0.1463	-10.6265	-1.8314	-1.2641	$<.001^{***}$
	Strident identity	-0.8427	0.1906	-4.4205	-1.2156	-0.4716	$<.001^{***}$
	Anterior identity	1.3207	0.5778	2.2855	0.1535	2.4554	0.016^{*}
	Coronal identity	-0.8464	0.2171	-3.8995	-1.2629	-0.4309	$<.001^{***}$
	Labial identity	-2.1668	0.3002	-7.2173	-2.7372	-1.5935	$<.001^{***}$
	Nasal identity	-0.0815	0.2503	-0.3257	-0.5714	0.4146	0.73
C_3	**Total identity**	-0.7889	0.1881	-4.1947	-1.1574	-0.4198	$<.001^{***}$
	Sonorant identity	1.5489	0.2462	6.2909	1.0774	2.0444	$<.001^{***}$
	Voice identity	-0.6932	0.2390	-2.9008	-1.1736	-0.2361	0.002^{**}
	Continuant identity	0.8278	0.2077	3.9866	0.4211	1.2334	$<.001^{***}$
	Strident identity	-2.3943	0.2254	-10.6203	-2.8504	-1.9424	$<.001^{***}$
	Anterior identity	0.2458	0.4789	0.5131	-0.7456	1.2054	0.618
	Coronal identity	0.4544	0.1658	2.7407	0.1197	0.7935	0.012^{*}
	Labial identity	-0.7807	0.2630	-2.9688	-1.3117	-0.2481	0.006^{**}
	Nasal identity	-1.1454	0.2120	-5.4039	-1.5793	-0.7288	$<.001^{***}$
C_4	**Total identity**	-0.9657	0.2389	-4.0422	-1.4505	-0.4703	$<.001^{***}$
	Sonorant identity	1.9742	0.2551	7.7381	1.4611	2.4885	$<.001^{***}$
	Voice identity	-0.6780	0.2347	-2.8890	-1.1401	-0.2026	0.004^{**}
	Continuant identity	0.1308	0.2288	0.5718	-0.3241	0.6022	0.616
	Strident identity	-3.4709	0.2483	-13.9762	-3.9699	-2.9808	$<.001^{***}$
	Anterior identity	-0.1093	0.4723	-0.2314	-1.0578	0.8355	0.806
	Coronal identity	0.2520	0.1974	1.2768	-0.1212	0.6165	0.184
	Labial identity	-0.3079	0.2828	-1.0886	-0.8769	0.2499	0.296
	Nasal identity	-1.7796	0.2056	-8.6571	-2.1915	-1.3756	$<.001^{***}$
	Transitional phonotactic probability	0.4553	0.0803	5.6728	0.3006	0.6109	$<.001^{***}$

Number of observations: 4,900; number of participants: 207; number of base forms: 30; number of word shapes: 4; number of linking consonants: 4
Level of significance: · ($p \leqslant 0.1$), * ($p \leqslant 0.05$), ** ($p \leqslant 0.01$), *** ($p \leqslant 0.001$).

3.2. Study II: Locality: distance-based decay and syllable role

In the following sections, we examine which consonants play a role in the identity avoidance effect and whether a distance-based decay effect can be found through a series of model comparisons in Section 3.2.1. We then repeat the same analyses with the five most frequent word shapes in our dataset in Section 3.2.2 to examine the role of syllable position in the sense of Bennett 2012; Rose and Walker 2004. The regression tables of all of the models can be found on the `osf.io` repository (see Section 5 *Data accessibility statement*).

3.2.1. *Identity avoidance beyond C_2*

To examine whether the identity avoidance effect operates beyond C_2, model comparisons were performed by comparing a full model with identity variables from all consonants with models without any identity variables of a specific consonant. The drop in AIC was used as a measure of variable importance.

All ΔAIC values are above 2 (a typical significance threshold for AIC values), therefore the identity avoidance effect operates over every single consonant, including C_3 and C_4. This is a surprising finding since most of the previous studies did not find an effect from beyond C_2 (e.g., Kelepir 2000; Wedel 1999, 2000).

To examine whether there is a distance-based decay effect such that the importance of each consonant decreases as the distance from the linking consonant increases, we compare the relative level of importance across consonant positions. In the two-consonant group, C_2 (ΔAIC = 872.30) is more important than C_1 (ΔAIC = 803.71). In the three-consonant group, C_1 (ΔAIC = 1679.46) is the most important consonant, followed by C_2 (ΔAIC = 1201.62) and C_3 (ΔAIC = 549.03). In the three-consonant group, C_1 (ΔAIC = 1679.46) is the most important consonant, followed by C_2 (ΔAIC = 1201.62) which in turn is more important than C_3 (ΔAIC = 549.03). In the four-consonant group, the consonants from the most important to the least are C_1 (ΔAIC = 628.49), C_2 (ΔAIC = 379.64), C_4 (ΔAIC = 370.35) and C_3 (ΔAIC = 179.20). The ΔAICs of C_3 and C_4 are lower than those of C_2 and C_1. However, the exact rankings do not clearly suggest a distance-based decay effect. While a distance-based decay effect can be seen with the three-consonant group ($C_1 > C_2 > C_3$), the order of importance diverges with the other two groups – 1) C_4 of the four-consonant group is more important C_3, and 2) C_2 of the two-consonant group is more important than C_1. We speculate that the divergence is attributable to the syllable position of the consonants. In the next section we examine the effect of syllable position by separating the items by word shape.[20]

3.2.2. *Syllable Position*

The method of evaluating variable importance of each consonant remains the same as section 3.2.1. The difference is that in Section 3.2.1, the data was divided up by the number of consonants the items contain, while in the current section, the data was divided up by the five word shapes to enable an examination of the syllable position. These five word shapes are C_1VC_2V, C_1VC_2, $C_1VC_2C_3V$, $C_1VC_2VC_3$ and $C_1VC_2C_3VC_4$. The word shapes with the same number of consonants were compared. C_1VC_2 and C_1VC_2V were compared because C_2 is a coda consonant in C_1VC_2 but an onset consonant in C_1VC_2V. Similarly, $C_1VC_2C_3V$ and $C_1VC_2VC_3$ were compared because C_2 and C_3 are syllabified different across the two shapes. Finally, $C_1VC_2C_3VC_4$ was examined because it could reveal the combined effect of syllable position and the distance decay effect, since it contains two sets of onsets (C_1 and C_3) and codas (C_2 and C_4) which differ only in their distance from the linking consonant.

AIC is being used only to compare models with the same set of data. For example, a fully-specified model was fitted over the subset of the data with the C_1VC_2V items. To

[20]We acknowledge that one could test the syllable position effect more directly by comparing two models, one with identity avoidance predictors and make references to the syllable constituency of their corresponding consonant, and one without making such references. We will leave such alternative analyses for the future.

evaluate the importance of C_1, we fitted a model on the same subset of the data but without any identity variables of C_1, i,e, the model regression structure has the identity variables of C_2 and not of C_1. AIC was computed for both of these models (the initial model with both identity variables of C_1 and C_2, and the new model with only identity variables of C_2). The two AIC values were used to compare these two models.

The word shapes with the same number of consonants were compared and visualised. Figure 1 visualises the variable importance of each consonant in C_1VC_2V and C_1VC_2 base forms. C_1VC_2V shows a distance-based decay effect with C_2 being less important than C_1, while the C_1VC_2 shows the opposite pattern with C_1 being less important than C_2. These differences indicate that the divergence observed in the two-consonant group (as described in Section 3.2.1) was driven by C_1VC_2 which has slightly more base forms than C_1VC_2V (23 C_1VC_2 base forms compared to 19 C_1VC_2V base forms). The primary difference between C_1VC_2 and C_1VC_2V is that C_2 is a coda consonant in C_1VC_2 but an onset consonant in C_1VC_2V. The extra vowel/syllable in C_1VC_2V is also a possible cause of the divergence. To further evaluate these observations, we turn to the two word shapes in the three-consonant group.

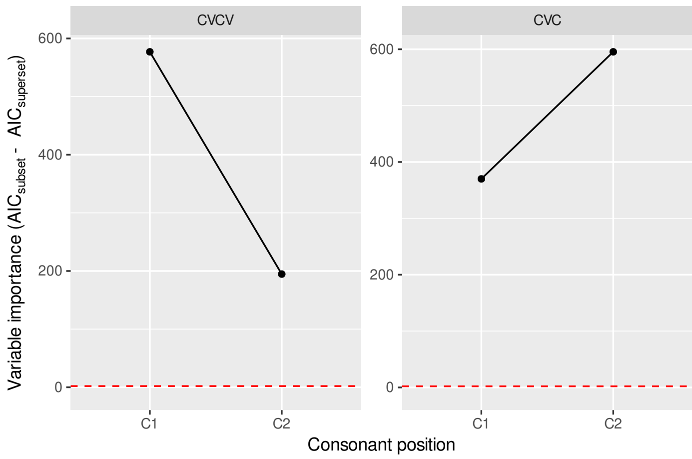

Figure 1. Variable importance of each consonant in two-consonant base forms by word shape: C_1VC_2V and C_1VC_2

Figure 2 visualises the variable importance of each consonant in $C_1VC_2VC_3$ and $C_1VC_2C_3V$ base forms. $C_1VC_2VC_3$ and $C_1VC_2C_3V$ both have three consonants and two syllables but they differ in terms of the constituency of their C_2 and C_3. The C_2 is an onset in $C_1VC_2VC_3$, but a coda in $C_1VC_2C_3V$. The C_3 is a coda in $C_1VC_2VC_3$, but an onset in $C_1VC_2C_3V$. $C_1VC_2VC_3$ shows a distance-based decay effect with a decrease in importance from C_1 to C_3. However, $C_1VC_2C_3V$ shows a different pattern. A general distance-based decay can still be seen with C_1 being more important than C_3, but C_2 diverges from the pattern being more important than both C_1 and C_3. $C_1VC_2C_3V$ matches the distance-based decay pattern observed in the three-consonant group (as described in Section 3.2.1). This is again not surprising since $C_1VC_2VC_3$ is the dominant word shape in the three-consonant group with 37 base forms, while $C_1VC_2C_3V$ has 14 base forms.

One explanation of the patterns of these four word shapes is to consider two effects operating in tandem. The first effect is the distance-based decay effect which predicts a decrease in importance as distance increases. The second effect is that the syllable position effect which predicts coda consonants to be more important than onset consonants. The pattern of $C_1VC_2C_3V$ can be explained if we consider both effects together. C_1 is more important than C_3 because of the distance-based decay effect, while the divergence of C_2

being more important than C_1 is because C_2 is a coda consonant. The C_2 of $C_1VC_2VC_3$ conforms to the distance-based decay effect and did not diverge because it is not a coda. Similarly, the divergence of C_2 from the distance-based decay effect in C_1VC_2 but not in C_1VC_2V can be explained since the C_2 in C_1VC_2 is a coda but an onset in C_2VC_2V. To further evaluate the two hypothesised effects, we turn to the four-consonant group.

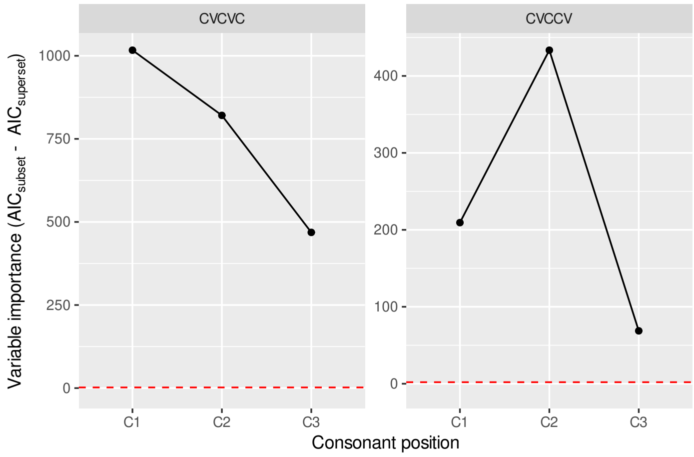

Figure 2. Variable importance of each consonant in three-consonant base forms by word shape: $C_1VC_2VC_3$ and $C_1VC_2C_3V$

Figure 3 visualises the variable importance of each consonant in $C_1VC_2C_3VC_4$ base forms. C_1 and C_3 are both onsets, while C_2 and C_4 are both codas. A distance-based decay effect can be observed with the two onsets and the two codas; C_1 is more important than C_3, and C_2 is more important than C_4. The syllable position effect can also be observed with C_2 and C_4 (codas) being more important than C_1 and C_3 (onsets).

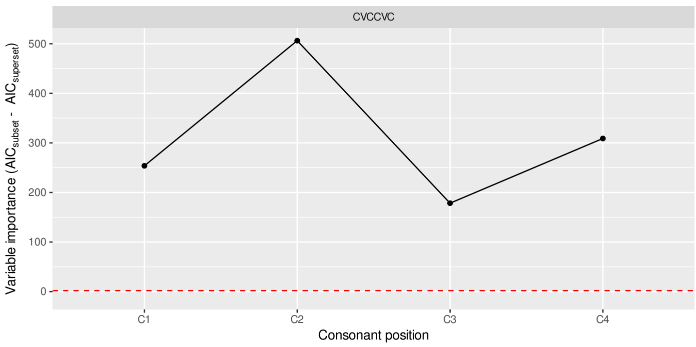

Figure 3. Variable importance of each consonant in four-consonant base forms by word shape: $C_1VC_2C_3VC_4$

4. Discussion

This section presents our results for the two studies. We discuss and situate our findings of the Turkish case into the broader literature, focusing on feature specificity and locality. The section also connects our findings to higher-level topics such as factors beyond identity avoidance and representation of speakers' knowledge.[21]

[21] See Appendix G for the discussion of the findings regarding the Turkish specific preference hierarchy relating to the linking consonant.

4.1. Feature specificity

The model comparisons in Study I revealed two key findings. The first finding concerns the importance of total identity and partial identity. While previous studies have evaluated both types of identity in their formal analyses, it is entirely conceivable that total identity might not be needed after partial identity has been taken into account. Partial identity is a function of individual featural identity. If two segments are totally identical, both total identity and partial identity would be able to capture the same degree of similarity. Given that the consonants in the base forms might not be one of the four linking consonants, partial identity is expected to be much more important factor than total identity. In other words, partial identity can in theory capture what total identity can but not vice-versa. Our result suggests that partial identity is more important than total identity across all three item groups but total identity still contributes above and beyond partial identity.

The second finding concerns the nature of partial identity. Recall that cross-linguistic studies have found that features that participate in identity avoidance processes are not weighed equally, in that they may differ in the extent to which they influence the phenomenon in question. Gallagher and Coon (2009) found that features such as [+strident] and [+ejective] trigger greater OCP effects than others in Chol. Moreover, non-coronal place features have been shown to trigger OCP effects of varying strengths (see e.g., Frisch et al. (2004); Pierrehumbert (1993)). In light of this background, we addressed the question of whether or not the individual features contribute equally in Turkish. It was found that the models with individual featural identity variables consistently outperformed the models with sum featural identity variables, therefore partial identity operates on an individual featural level. This finding echoes many of the formal analyses from previous studies which formulated a number of identity-avoidance constraints using specific features. In particular, Demircan (1987) identified coronal, labial and nasal as features of importance and Kelepir (2000)'s analyses involve strident, labial, continuant and sonorant.

To determine whether these six features are of particular importance in Turkish partial reduplication in general, we inspect the model summary of each of the three item groups in Tables 4, 5, and 6[22]. Strident appears to be the most prevalent feature since it was significant in all of the consonant positions across all three models. Labial was similarly prevalent since it was only insignificant once (C_4 of the four-consonant model). Nasal was insignificant twice (C_2 of the two-consonant model and the four-consonant model). Continuant was insignificant three times (C_1 of the two-consonant model, and C_1 and C_4 of the four-consonant model) and it was in the opposite direction of identity avoidance twice (C_3 of the three-consonant and the four-consonant model). Sonorant was insignificant three times (C_1 of the three-consonant model and the four-consonant model and C_2 of the three-consonant model) and it was in the opposite direction in C_3 of the four-consonant model. Coronal was insignificant four times (C_1 of the three-consonant model and the four-consonant model, C_2 of the two-consonant model and C_4 of the four-consonant model) and it was in the opposite direction in C_3 of the four-consonant model. In terms of the size of the coefficients, strident and labial were in generally higher than the other features. The features, continuant, sonorant and coronal, were not as consistent in terms of their statistical significance and the direction of their effects.

Overall, these models suggest that strident and labial were the most prevalent features. This finding is in line both with two of the previous studies on Turkish, i.e., Kelepir 1999; Yu 1999, which emphasised the importance of [strident], as well as cross-linguistic studies which found a similar effect (e.g., Bye 2011; Gallagher and Coon 2009). The feature [labial] has also been found to be influential as opposed to some other features, such as [nasal],

[22]Note that we are not comparing the effect sizes of the features (the coefficients of the features) across models.

which again are in line with the cross-linguistic typology, in that non-coronal place features were argued to play a larger role in the OCP literature (e.g., Bye 2011; Pierrehumbert 1993). Another finding of our study concerns the status of [coronal]: while both Demircan (1987) and Kelepir (1999) argue for the importance of coronal feature in their systems, our results found no such effect. In this regard, our finding is more consistent with the cross-linguistic generalization that coronal is not an influential feature (e.g., Bye 2011; Coetzee and Pater 2008; Pierrehumbert 1993).

While our findings are in line with cross-linguistic typology, it is important to remind ourselves that these cross-linguistic tendencies might be triggered by language-specific phenomena. In the case of Turkish, the alternating segments being [p, m, s, r] largely determine which features could *potentially* participate in identity avoidance. As such, although [strident] is influential in both Chol and Turkish, this importance is due to different factors.

We also believe that it is unsurprising that strident, labial and nasal are particularly prevalent. Strident and nasal decrease the preference for [s] and [m] respectively, while labial decreases the preference for [p] and [m]. The three features together influence the preference for three of the four linking consonants except for [r]. The preference for [r] can be captured by using a markedness constraint since, as we have discussed earlier, it is the least preferred linking consonant even when other fixed and random effects were taken into account. In contrast to strident, for example, continuant was not a prevalent feature in our study, contrary to Kelepir's analyses. Similarly, the nasal feature is important than the sonorant feature, also used in the previous literature. These findings, we believe, further confirm the heuristic side of the inventory of features employed by many previous studies. Additionally, we speculate that this state of affairs could be explained by resorting subset-superset relation between the features in question. In particular, the feature sonorant encompasses a larger set of consonants as opposed to nasals; likewise, continuant picks out a larger set of consonant than strident. The results indicate that the feature that is more specific applies first or is more influential, in a way that makes the superset feature redundant.

Our results also reveal that the identity effects of a given feature for a given consonant can differ across different stem types. For example, the continuant identity effect of C_1 is not significant in bi-consonantal and quadri-consonantal stems, but it is significant in tri-consonantal stems.

These apparent inconsistencies call for some *post-hoc* speculative remarks. While it is plausible to assume that the same feature in the same position will exert the same strength across the board for stems of different lengths, we believe this does not need to be the case. The strength of the effect of a feature may very well be not determined once and for all, rather it may be contextually determined, which might include the number of other consonants and the syllable roles in the stem. For example, while a feature F in C_1 might be strong in a two-consonant stem, the same feature F might have a weaker role in a three-consonant stem if a competing feature is found in C_3 position. In statistical terms, additional consonants could introduce additional identity avoidance effects which could influence the significance of other consonants (e.g., C_1) by taking up variance in the regression model. As an analogy, in phonological theories like OT, constraints are not strictly-ranked, but rather are weighted, where multiple violations of lower-priority constraints are able to overcome the violation of a higher-priority constraint (the "gang effects") (Pater 2016). As such, we should not necessarily interpret the mismatches/inconsistencies about the effect of a particular feature as a negative or as a shortcoming of the model, but rather as indicators of the contextual factors that give rise to the apparent mismatch.

Relatedly, what can our statistical models, which were fitted separately over different word shapes, tell us about the Turkish speakers' grammar? The models were trained on

Turkish speakers' responses to real words which as discussed in Section 2.2, are confined to certain feature combinations. Therefore, our models would allow us to predict how a Turkish speaker would choose C_2 when emphatically reduplicating a novel existing or nonce stem of given length with a C with a given feature specification in a particular position, as long as the stems have comparable feature combinations as the real words that we examined.

If our set of real words are generally representative of the reduplicable items in the Turkish lexicon, then two theoretical stances are conceivable in terms of what aspects of speakers' grammar our models are capturing. On the one hand, one could take a position with a usage-based lexicon-based approach (e.g., Baayen, Chuang, Shafaei-Bajestan, Blevins, et al., 2019; Bybee, 2003, 2006, 2010; Chuang & Baayen, 2021). Under this view, Turkish speakers' grammar is confined to the attested feature combinations in their mental lexicon. Therefore, there are limits on what a Turkish speaker can learn and thus predict from the stems they know, and therefore on what we, as the modelers, can learn and thus predict from our models of what a Turkish speaker can learn from their lexicons. On the other hand, one could take a different position which allows for Turkish speakers' grammar to not be solely based on their mental lexicon, but rather also under the influence of learning strategies, such as analogical learning (e.g., Arndt-Lappe, 2014; Nosofsky, 1986; Skousen, Lonsdale, & Parkinson, 2002) and discriminative learning (e.g., Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011). Further research is needed to tease these two stances apart by examining nonce words that contains unattested feature combinations.[23,24]

4.2. Locality: distance-based decay and syllable role

The analyses of locality revealed a number of findings, including the domain over which this phenomenon operates, whether its effect is weighed equally across the domain or other factors such as syllable position effects are at play.

The first finding concerns the rightward boundary up to which the identity avoidance effect applies. Recall that most previous studies emphasise the importance of C_1 and C_2 (e.g., Kelepir 1999; Wedel 1999), with some of them (explicitly) assuming a cut-off after C_2 (Wedel 1999). At the same time, Yu (1999) found that the strident feature has an effect with respect to C_3, which suggests that the LC is not restricted to identity avoidance effect only with respect to C_1 and C_2. Our results reveal that Yu (1999)'s conclusion was on the right track, but goes further to include all the consonants in the base, and that the effects of the other consonants in the base were not particularly strong or insignificant compared to C_1 and C_2.

The second finding shows that despite all the consonants contributing to the identity avoidance effect, the effect was not uniform but linear, with the strength of the effect decreasing further into the base. This suggests all the consonants play a role in identity avoidance but they are subject to a distance-based decay effect (Zymet, 2014, 2018), which states

[23]We thank John Kingston and Gaja Jarosz for extensive discussions on this topic.

[24]Another possibility is that our sample of real words are not representative enough of the Turkish lexicon and thus reflecting only an experimental accident with our item selection. In other words, our apparent inconsistent effects could be the result of how our model overfitted the sampled data. Future research could conduct a lexical analysis of how representative our items are, and if they were unrepresentative then one could extend our experiments to a larger set of real words.

Furthermore, yet another possibility is that we have not taken into account of factors beyond the identity avoidance effect as will be discussed in Section 4.3, as well as factors that we excluded in this study (as described in Appendix B). Future research could incorporate these additional factors to examine whether and how they would influence our complete findings, including both the apparent inconsistent effects and the consistent effects.

that the likelihood for the application of a phonological process decreases as transparent distance increases.

The third finding is that the distance-based decay effect is not completely linear, and is shown to be sensitive to the word shape. Specifically, some consonants diverged from the linear order, with C_3 of the four-consonant base forms is more important C_4, and C_2 of the two-consonant base forms is more important than C_1. We found that these divergences can be fully explained, once the constituency of a consonant is taken into account. We postulated that on top of the distance-based decay effect, the syllable position effect is also present with a coda consonant exhibiting a stronger identity avoidance effect than an onset consonant. When considering both effects in tandem, we were able to capture the patterns of importance across a number of frequent word shapes in our dataset. To further illustrate the syllable position effect, we can consider a minimal pair from our dataset – *sı**k*** 'tight' and *sı**k**ı* 'frequent'. In both base forms, the C_2 /k/ should disprefer the linking consonant [p]. Given the syllable position effect, C_2 /k/ in *sı**k*** should disprefer [p] more than that in *sı**k**ı*. This is indeed confirmed by our rating study. *sıp* + *sık* has a mean rating of 3.7 which is less acceptable than *sıp* + *sıkı* which has a mean rating of 5.3. The coda effect was observed in a recent nonce word study. Köylü (2020) asked 14 native speakers of Turkish to reduplicate 48 nonce-words of four word shapes (VCV, CVC, CVCV, and VCCV). It was found that the linking consonant was never identical to C_1 or C_2 with the CVC nonce-words. However, the linking consonant was never identical to C_1 but sometimes identical to C_2 with the CVCV nonce-words. In other words, C_2 did not always affect the linking consonant in terms of total identity in CVCV but not in CVC. This can be explained by how C_2 is a coda in CVC which has a stronger effect on the linking consonant than when it is an onset in CVCV. Why should a coda consonant outweigh an onset consonant? We speculate that it is due to the fact that the linking consonant itself is also a coda consonant for consonant-initial base forms. Since the identity avoidance effect is a function of the similarity between a consonant in the base and the linking consonant, the similarity of the two consonants would be stronger if they have the same type of constituency. This finding lends support to the view that syllable position might play a role in contributing to segments' (dis)similarity (Bennett 2012; Rose and Walker 2004).[25] In particular, the effect would decay with the distance from the linking consonant and would be enhanced if the consonant in the base matches the constituency of the linking consonant (coda vs onset). By uncovering the presence of this effect in the Turkish emphatic reduplication, our study also adds to the typology of syllable position effects. As mentioned earlier in the paper, Bennett (2013) notes that while harmony/assimilation is predicted between consonants with matching syllable roles, dissimilation is predicted for consonants with mismatching syllable roles (but not those with matching syllable roles). Turkish constitutes an example in which dissimilation favors matching syllable roles, with the effect of syllable role observed in non-categorical patterns with multiple features.

All in all, Turkish serves as a fruitful testing ground in showing that the identity avoidance effect holds for all the segments, with the effect being strongest from C_1 and being weakened as a function of its distance from the target segment. Moreover, our study reveals that in addition to the distance, the syllable position also plays a role in the application of the partial reduplication. Given our findings, one could formulate the grammar by deriving the weights of the distance-decay function (e.g., the decay parameter of the decay function which describes the shape of the decay), the weights of syllable position (e.g., the weight of coda), and their possible interactions (McMullin & Burness, 2021; Zymet, 2014). This,

[25]It is possible to recast the empirical findings from our study in more formal terms such as Structural Surface Correspondence constraints (e.g., Structural SCorr-CC, which would limit correspondence to only those consonants which have matching syllable roles) as proposed by Bennett (2012) and Rose and Walker (2004).

however, will likely require a more well-balanced dataset with nonce words in terms of the representation of particular features, word shapes and the number of consonants in the base. This is left for future studies.

4.3. Beyond identity avoidance

The current study examined the Turkish partial reduplication phenomenon as a phonological operation with a focus on the identity avoidance effect. Our models were able to explain a sizable portion of the variance of our naturalness judgements[26]. The identity avoidance factors captured around 25% to 35% of the variance (two-consonant group: 26.96%, three-consonant group: 33.67%, and four-consonant group: 28.36%) and, together with the random effects which captured the idiosyncracies of lexical items, the models were able to explain around 50% to 70% of the total variance (two-consonant group: 55.73%, three-consonant group: 55.75%, and four-consonant group: 70.98%). Our study therefore provided ample evidence that phonological factors play a major role in the process behind Turkish partial reduplication. However, there is still variance to be explained by factors beyond phonology, such as the lexicon, morphology and semantics.

As also noted by an anonymous reviewer, our analysis captures a large portion of the patterns, yet certain minimal contrasts between reduplicated forms such as *köp-kötü* 'very bad' vs. *kas-katı* 'very hard' still are not fully accounted for. As shown in the Appendix, different linking consonants are selected for the same *k-t* sequence of consonants in the base, and in the same syllabic positions (as argued in Yu 1999, we take it that vowels do not play a role in the reduplication process). Previous analyses have also failed in providing a satisfactory explanation for such minimal contrasts. For example, Demircan (1987) brings up the intuition that speakers might avoid a reduplicant that resembles an existing root and confirms it in a corpus study by Kılıç and Bozşahin (2013), which demonstrated that root-level lexical statistics inversely correlate with the preference of a linking consonant. Speakers' preference therefore might depend on the knowledge of distributional statistics at the morpholexical level. Speakers might not prefer to use a reduplicated form for the emphatic meaning, but instead prefer to use the word *çok* 'very' with the base form. However, in the case of *kas-katı*, it is still unclear why *s* is selected over *p* since both *kas* 'muscle' and *kap* 'container' are meaningful words in Turkish. In fact, the only linking consonant that does not result in a form resembling another Turkish root is *m* as *kam* is not a Turkish word, whereas *kar* 'snow' also is). Therefore, we acknowledge the presence of factors beyond phonological considerations that play a role in the choice of an LC.[27]

In a series of corpus and experimental studies by Kaufman (2014), it was observed that the preference to reduplicate a base form depends on its semantic class and the semantic class of existing base forms that are frequently reduplicated. A low rating of a reduplicated form might not be due to phonological factors but rather due to the participant's dispreference to reduplicate the base form. Incorporating morpholexical statistics and semantics can therefore provide a more complete picture of the Turkish partial reduplication phenomenon and we leave this for future research.

In light of the phonological generalizations uncovered by our experimental study, in the

[26]The proportion of variance captured by fixed effects in the models was computed with the function r.squaredGLMM(), part of the MuMIn library in R (Bartoń, n.d.). This function returns both marginal R^2 and conditional R^2. Marginal R^2 represents the variance explained by fixed factors. Conditional R^2 represents the variance explained by fixed and random factors.

[27]The same anonymous reviewer also notes another issue for the pair *dip-diri* vs. *dup-duru*. Both have the same consonant sequence, the same word shape, and the same linking consonant. However, the acceptability ratings are found to be different for *dip-diri* (mean: 1.43) vs. *dup-duru* (mean: 1.35). This calls for a careful examination of other factors, such as frequency of forms in a corpus of Turkish in future studies.

next section we discuss what speakers' knowledge of this pattern looks like, i.e., what is the potential representation speakers have in mind when using this phenomenon?

4.4. Representation of speakers' knowledge

With respect to the representation of speakers' knowledge of the Turkish partial reduplication phenomenon, one point of investigation that researchers have focused on is whether the choice of a particular linking consonant in partial reduplication is simply a matter of *lexicalization*, a term which has been interpreted in more than one way in the literature. One interpretation of this approach is whether the choice of the LC is random/arbitrary, or follows a set of generalizations or rules. If the conclusion is that the choice of the LC is arbitrary, then the phenomenon is considered to be lexicalised. Although earlier studies assumed the choice of the LC to be arbitrary, thus lexicalised (e.g., Foster 1969; Lewis 1967; Yavaş 1980), a number of studies have argued that the choice is not lexicalised, and is indeed conditioned by various rules. These studies include Hatiboğlu (1973), Demircan (1987), Dobrovolsky (1987), Taneri (1990), Wedel (1999), Yu (1999), Kelepir (2000), Sofu and Altan (2008), Kaufman (2014). While these works vary considerably in their implementations of the observations, they converge on the view that the choice of the LC is not arbitrary or lexicalised, and that it is subject to several dissimilation constraints motivated by the OCP, similar to the analyses given for dissimilation processes in other, unrelated languages.

Some other studies have used nonce-words as a diagnostic as to whether Turkish partial reduplication is lexicalised or productive. Under the (often implicit) assumption that real words that participate in this phenomenon obey various rules or generalizations, studies in this line of research probe whether nonce-words are subject to the same generalizations. It turns out the results from these studies are far from clear, and have found varying, and sometimes conflicting results. For example, while Sofu (2005) concludes that speakers seem to extend at least some of the rules to nonce-words, two more recent studies, Demir (2018) and Köylü (2020), reach opposing conclusions as to the status of lexicalization and productivity of the reduplication patterns. While Demir (2018) interprets her results in favour of a lexicalization approach, Köylü (2020) argues that speakers do extend the same strategies they use for real words to nonce-words.

On the side of studies that investigate whether the Turkish emphatic reduplication is arbitrary or rule-governed, Wedel (1999) for example, argues that 'native speakers do abstract some productive phonological generalization from the emphatic forms that exist', as such speakers 'have access to a uniform, constraint-based schema' in using this phenomenon. This is at least the case for the LCs [p, s, m], while the LC [r] may indeed have lost its productivity since it is not used with novel forms and is the least utilised LC as confirmed by other studies including Yu (1999) and this current study.[28] Along the same lines, Yu (1998) also notes that 'modern speakers of Turkish have some grammatical knowledge of emphatic reduplication' on the basis of the results that speakers do not blindly choose a particular LC, e.g., [p], across the board, and concludes:

> Proposals that claim the emphatic construction is unproductive and that all reduplicative closers must be lexically listed with the base form must be taken with great precaution, if not rejected altogether. The experiments reported here clearly suggest that native speakers if Turkish still retain some grammatical knowledge

[28]An anonymous reviewer suggested that for Wedel (1999), all reduplicated forms are lexicalised since they are not productive. Given the above statements from Wedel (1999), we believe (see also Köylü 2020 for this view) that Wedel does not take this phenomenon to be lexicalised, differing from the interpretation of that study by the reviewer.

> of the selectional restriction of the closer ill the emphatic reduplication construction. (Yu 1998:39)

In this regard, we also add that if partial reduplication was simply a matter of lexicalization, it would be surprising that for a major number of items, speakers exhibit variation, in the sense that more than one linking consonant is permitted. Crucially, in these variable cases, the permissible linking consonants are not identical across items, and even for items that permit multiple identical LCs, the relative acceptability differs from one to another and across speakers. All these suggest that speakers are not blindly memorizing a specific LC for each base form, but instead, as Wedel (1999) and Yu (1999) already suggested, observe certain phonological generalizations, which our study aims to make precise.

Let us now turn to the other interpretation of lexicalization, which focuses on whether nonce-words parallel the behavior of real words. For example, a study by Demir (2018) examines the choice of LC with adults via a comparison of 10 real- and 34 nonce-words using an open-set response task. In one experiment, nonce-words are chosen to have a real word counterpart with identical consonants and word shapes, but different vowels. Demir finds that real words follow the expected observations noted by previous studies, as such using the four common LCs, [p,m,s,r]. On the other hand, in the case of nonce-words, speakers resort to strategies that are not available for real words. For example, they might copy the CVC from the base, as opposed to the CV plus LC strategy, or they might omit a linker altogether. Moreover, nonce-words which were chosen to analogise the real words do not show the same linking consonant as their supposed real word counterparts. Demir (2018) interprets these divergent results between real- and nonce-words to mean that there is not a pattern or set of generalizations for emphatic reduplication of real words, which one would expect to be extended to nonce-words. Therefore, reduplication of real words must be lexicalised.

However, this conclusion might be a bit too hasty. As we just noted, there are other nonce-word studies carried out on Turkish partial reduplication that arrive at different conclusions or interpretations. For example, Sofu (2005) examines the choice of LC with both adults and children using 38 nonce-words. She finds that adults and children conform to the expected patterns in their use of the classic LCs, [p,m,s,r], while children use [p] more than adults. Moreover, children use linking consonants that are different from [p,m,s,r] much more often than adults, e.g., [t,n,f]. This study shows that both adults and children do extend the patterns observed in real words to nonce-words as well.[29] Similarly, a more recent study by Köylü (2020) also concludes that Turkish native speakers extend the reduplication strategies they employ in real words to nonce-words. These conflicting results call for a careful investigation regarding the causes behind them. Here we speculate on a few potential issues.

There might be various methodological or linguistic factors that lead to this divergence in nonce-word studies. For example, all the previous studies simply present the nonce-word test items by themselves out of a context without assigning them any meaning. This is not a trivial choice in light of the fact that reduplication is applicable only to gradable modifiers. It is likely that these out-of-context nonce-words were not interpreted as such by the participants, and therefore participants resorted to strategies that differ from real words, whose meaning and property of being gradable they are aware of. The restriction regarding the category and property this reduplication requires could also be a factor as to why nonce-word studies might not be replicating or reaching lower scores.

[29]Regarding the use of [p] more frequently and the presence of non-standard linking consonants, it could be showing that children are still in the process of mastering the abstract generalizations. As such, they sometimes revert to the default LC, p, or have a larger set of potential LCs that they have not narrowed down yet to the ones adults use.

This last point also relates to another concept, productivity, which usually comes up in the discussion of lexicalization. On the point of productivity, it is worth highlighting that it should be approached with caution. This is because emphatic reduplication, as just noted above, applies specifically to a subset of adjectives i.e., gradable adjectives, (and adverbs, which usually are built on adjectives in Turkish) and not to absolute adjectives or modifiers in general. As such, we can make sense of why it is found in a relatively small number of items in the language, due to its nature. Therefore, any statement about productivity should take this important aspect into consideration, and its potential role especially on studies investigating nonce-words.

Moreover, recall that Demir (2018) created certain nonce-words in anticipation of analogy with real words. For example, the nonce-word *mava* was created by Demir on the assumption that the participants would analogise it to the existing adjective *mavi* 'blue' while they are attempting to reduplicate *mava*. This is not an innocuous assumption, however. First, analogy does not rely on just a single word, but a number of words (e.g., the Generalised Neighbourhood Model by Bailey and Hahn (2001) considers all the words in the lexicon weighted by lexical frequency and form similarity to the target (nonce-)words). As such this assumption overlooks the complicated aspect of how analogy works. Secondly, some words might have a large number of noun neighbors, which might make the reduplication harder. For example, in the context of the nonce-word *boyuz* from Demir's study, our native intuition (and those of our consultants) analogises *boyuz* to *boyoz*, which is an existing word that refers to a food item. As such, although the experimenter might have a certain real word in mind while designing the nonce-words, participants might have completely different real word analogies. Similarly, with *mava*, native speakers we have consulted brought up the nouns *hava* 'air', or *tava* 'pan' as the first words that come to their minds, both of which are nouns, rather than the adjective *mavi* that Demir had in mind as a control item. This is significant since the control items that Demir has in mind may not be the items participants are supposedly analogizing to, as such might explain part of the results. A further related note is that analogy may not be solely based on phonological properties, but might be due to semantic resemblance participants establish with real words they might think of.

As just discussed in Section 4.3, there are also other factors beyond identity avoidance, such as lexical (Kılıç & Bozşahin, 2013) and semantic factors (Kaufman, 2014), that play a role but were not consistently considered in these nonce-word studies.

In light of these considerations, the results of our study are in support of the view that partial reduplication is subject to various active phonological rules, particularly speakers have access to locality and feature-based conditions (or generalizations) that they are applying to items that are potentially intensifiable. In this regard, it corroborates the findings/intuitions raised by studies such as Wedel (1999); Yu (1998) and accords with Demircan (1987) and Kelepir (1999) who also conclude that speakers obey various phonological rules when they do partial reduplication. In particular, in this study we have uncovered that the phonological rules that are exhibited by the real words are much more graded than previously thought. The identity avoidance effect is both locality-sensitive (distance-based decay effect, and syllable position effect) and feature-sensitive (individual features with different weights).

With that said, it is important to keep in mind that as Frisch et al. (2004) argues, lexical information, including lexical idiosyncrasies, and rules are not mutually exclusive, in that there does not need to be a categorical choice between the two interpretations. It is possible to have a phenomenon which respects various constraints, e.g., locality- or feature-specificity, or phonotactic ones. As such, it is not necessarily the case that presence of phonological rules conflicts with or rules out the presence of lexical information. The Turkish emphatic reduplication may very well be an example of this sort. Wedel's (1999) conclusion might also be in line with this interpretation, in that while certain linking consonants [p, s, m] still actively participate in the identity avoidance effect, [r] might have fallen on the memorization side of it.

Having better understood the nature of the phonological grammar that speakers might have, this naturally leads to the second question about lexicalization, i.e., whether and to what extent these rules are being extended to nonce-words. Concerning nonce-words, we can bring in the insights raised by Becker, Ketrez, and Nevins (2011) who found, on the basis of another phenomenon in Turkish, that nonce-words are subject to only some of the rules exhibited by real words (see also Harris, Neasom, and Tang 2016; Hayes and White 2013). As such, there is no *a-priori* reason to expect that nonce-words fully conform to the same generalizations as real words, or reflect all of the generalizations/rules found for real words. Moreover, as discussed above, any study that aims to investigate nonce-words in the Turkish emphatic reduplication must also address a number of methodological and linguistic factors such as framing the nonce-words in a context that signifies its meaning and property of being gradable, as well as controlling for lexical and semantic factors. This can be done by using a combination of careful stimuli design and statistical modelling (Redington & Chater, 1996; Tang & Baer-Henney, 2023). Once these factors are controlled for, we expect (following Becker et al. 2011) that at least some of the real word generalisations would apply to nonce-words (some of which have already been argued to be the case, e.g., Köylü 2020; Sofu 2005).

5. Conclusions

This paper has re-examined a well-known reduplication phenomenon in Turkish. Modifiers such as adjectives and adverbs can undergo a partial reduplication process to express an emphatic meaning by prefixing a C_1VC_2 syllable. Unlike most instances of reduplication with fixed segmentism which have a single fixed segment, the Turkish emphatic reduplication contains four fixed segments, as such the linking consonant C_2 can be one of the four consonants: [p], [m], [s] and [r]. The study investigates the factors conditioning the choice of the LC, by focusing the nature of the (dis)similarity (feature specificity) and the proximity (locality) between the consonants in the base and the LC. Turkish emphatic reduplication turns out to be well-suited for shedding light on the nature of identity avoidance as it allows multiple possible fixed segments, and the same item itself might be used with multiple LCs.

Using an acceptability rating task conducted with over 200 participants for 162 base forms, the study has uncovered a number of significant findings with implications for the broader research on identity avoidance. Our analyses revealed that the identity avoidance effect is much more graded than it has been previously proposed in both its specificity and locality. Unlike most previous studies which emphasise the importance of C_1 and C_2 of the base as the domain in which dissimilation operates over, our study has found that the effect extends over all consonants in the base. Crucially, despite all the consonants contributing to the identity avoidance effect, the effect was not uniform, with the strength of the effect decreasing further into the base. Therefore, we demonstrate that the phenomenon is subject to a distance-based decay effect (Zymet, 2014, 2018). Moreover, the study uncovers an intricate interplay between the distance-based decay effect and the syllable position effect (Bennett 2012; Rose and Walker 2004), which is not as transparent as it is in most other languages. This novel finding for the Turkish emphatic reduplication is made possible due to the methodology and statistical tools adopted in this study.

In terms of the nature of specificity, our results reveal that in the Turkish emphatic reduplication process, the similarity between the consonants in the base and the LCs operates at the segmental level (total identity) as well as the level of individual phonological features (partial identity). Moreover, in line with the cross-linguistic picture, we have found that features that participate in identity avoidance processes may differ in the extent to which they influence the phenomenon in question. Our study confirms the importance of only some of the features employed in the previous studies, such as [strident], [labial] and [nasal], but not others, such as [coronal], [sonorant], or [continuant]. The important features like [strident] and [labial] and the unimportant features such as [coronal] and [sonorant] are more in line with the cross-linguistic tendencies (see e.g., Bye 2011; Pierrehumbert 1993). Methodologically we demonstrate that the precise nature of the identity avoidance effect can be revealed using hierarchical regression and statistical model comparisons (Graff & Jaeger, 2009; Zymet, 2019).

Data accessibility statement

Given the lack of rating judgements for the Turkish partial reduplication phenomenon, we made our data available in an Open Science Framework repository (`https://www.doi.org/10.17605/OSF.IO/P2JDK`) and in the Appendix H. Furthermore, given the complexity of the analyses used in this study, we made the analysis scripts we used to produce the results available so that readers can evaluate the data and our procedures themselves.

Abbreviations

The following abbreviations were used.

LC: Linking consonant

AIC: Akaike information criterion

BIC: Bayesian information criterion

Additional files

The appendix consists of nine parts A, B, C, D, E, F, G, and H. In part A, we provide the phonological feature values of Turkish consonants. In part B, we provide a set of excluded variables and the reasons of their exclusion. In part C, we provide the distribution of the variables (both the response variable and the predictors) for each of the three item groups with two, three and four consonants in the base form. In part D, we provide the pairwise association results between the response variable and each of the predictors in Section 2.3.1 for each of the three item groups. In part E, we provide the random effects summaries for all the reported models. In part F, we provide the model comparison for feature specificity. In part G, we provide a discussion of the findings regarding the Turkish specific preference hierarchy relating to the linking consonant. In part H, we provide the by-item acceptability ratings as well as an inter-rater reliability analysis.

Acknowledgements

We dedicate this paper to the residents of all the towns that were impacted by the 2023 earthquakes in southeastern Turkey. We thank our Turkish-speaking participants for their generous contribution to our research. We thank John Kingston, Gaja Jarosz, Ryan Bennett, Joe Pater, Andrew Nevins, Ryan Budnick, Akiva Bacovcin, John Harris, Fabian Tomaschek, Yılmaz Köylü and the audience at the 26th Manchester Phonology Meeting and the 15th Old

World Conference in Phonology for their feedback on earlier versions of the study. Special thanks to the anonymous *Laboratory Phonology* reviewers. All errors remain ours.

Competing interests

The authors have no competing interests to declare.

Authors' contributions

We follow the CRediT taxonomy[30].

Kevin Tang: Conceptualization, Methodology, Data curation, Formal analysis, Writing-Original draft preparation, Visualization, Investigation, Writing-Reviewing and Editing.

Faruk Akkuş: Conceptualization, Methodology, Data curation, Formal analysis, Writing-Original draft preparation, Visualization, Investigation, Writing-Reviewing and Editing.

References

Albright, A. (2007). *Gradient phonological acceptability as a grammatical effect.* Retrieved from http://web.mit.edu/albright/www/papers/Albright-GrammaticalGradience.pdf

Alderete, J., Beckman, J., Benua, L., Gnanadesikan, A., McCarthy, J., & Urbanczyk, S. (1999). Reduplication with fixed segmentism. *Linguistic Inquiry*, *30*(3), 327–364. doi: 10.1162/002438999554101

Arndt-Lappe, S. (2014). Analogy in suffix rivalry: The case of English-ity and-ness. *English Language & Linguistics*, *18*(3), 497–548. doi: 10.1017/S136067431400015X

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., Blevins, J. P., et al. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, *2019*. doi: 10.1155/2019/4895891

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–482. doi: 10.1037/a0023851

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591. doi: 10.1006/jmla.2000.2756

Bartoń, K. (n.d.). MuMIn: Multi-Model Inference [Computer software manual]. (R package version 1.43.6)

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Becker, M., Ketrez, N., & Nevins, A. (2011). The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language*, 84–125. doi: 10.1353/lan.2011.0016

Becker, M., & Levine, J. (2013). *Experigen: an online experiment platform.* Available at http://becker.phonologist.org/experigen.

[30] https://www.ucl.ac.uk/library/research-support/open-access/credit-taxonomy

Bennett, W. G. (2012). Dissimilation by correspondence in Sundanese. In N. Arnett & R. Bennett (Eds.), *Proceedings of the 30th West Coast Conference on Formal Linguistics* (pp. 76–86). Cascadilla Proceedings Project.

Bennett, W. G. (2013). *Dissimilation, Consonant Harmony, and Surface Correspondence* (Unpublished doctoral dissertation). Rutgers University.

Berkley, D. M. (2000). *Gradient obligatory contour principle effects* (Unpublished doctoral dissertation). Northwestern University.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. doi: 10.3758/BRM.41.4.977

Bybee, J. (2003). *Phonology and language use* (Vol. 94). Cambridge University Press.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711–733. doi: 10.1353/lan.2006.0186

Bybee, J. (2010). *Language, usage and cognition.* Cambridge University Press.

Bye, P. (2011). Dissimilation. In *The Blackwell Companion to Phonology* (p. 1-26). John Wiley & Sons, Ltd. doi: 10.1002/9781444335262.wbctp0060

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example.* John Wiley & Sons.

Chuang, Y.-Y., & Baayen, R. H. (2021, 12). Discriminative learning and the lexicon: NDL and LDL. In *Oxford Research Encyclopedia of Linguistics.* Oxford University Press. doi: 10.1093/acrefore/9780199384655.013.375

Clements, G. N., & Ridouane, R. (2006). Distinctive feature enhancement: a review. In A. Botinis (Ed.), *Proceedings of the ISCA tutorial and research workshop on experimental linguistics* (pp. 97–100). International Speech Communication Association.

Coetzee, A. W., & Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory, 26*(2), 289–337. doi: 10.1007/s11049-008-9039-z

Demir, N. (2018). Turkish reduplicative adjectives and adverbs. In P. Farrell (Ed.), *Proceedings of Linguistic Society of America* (Vol. 3, pp. 1–14). Linguistic Society of America. doi: 10.3765/plsa.v3i1.4300

Demircan, O. (1987). Emphatic reduplication in Turkish. In H. E. Boeschoten & L. T. Verhoeven (Eds.), *Studies on modern Turkish: Proceedings of the third conference on Turkish linguistics* (pp. 24–41). Tilburg University Press.

Dobrovolsky, M. (1987). Why CVC in Turkish reduplication. In P. Lilius & M. Saari (Eds.), *The Nordic languages and modern linguistics* (Vol. 6, pp. 131–146). Helsinki University Press.

Erguvanlı Taylan, E. (2015). *The phonology and morphology of Turkish.* İstanbul: Boğaziçi University Press.

Foster, J. F. (1969). *On some phonological rules of Turkish* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Frampton, J. (2009). *Distributed reduplication.* MIT Press.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory, 22*(1), 179–228. doi: 10.1023/B:NALA.0000005557.78535.3c

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language, 77*(1), 91–106. doi: 10.1353/lan.2001.0014

Gallagher, G., & Coon, J. (2009). Distinguishing total and partial identity: Evidence from chol. *Natural Language & Linguistic Theory, 27*(3), 545–582. doi: 10.1007/s11049-009-9075-3

Goldrick, M. (2011). Using psychological realism to advance phonological theory. In

The Handbook of Phonological Theory (p. 631-660). John Wiley & Sons, Ltd. doi: 10.1002/9781444343069.ch19
Graff, P., & Jaeger, T. (2009). Locality and feature specificity in ocp effects: Evidence from Aymara, Dutch, and Javanese. In M. R. Bochnak, P. Klecha, A. Lemieux, N. Nicola, J. Urban, & C. Weaver (Eds.), *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 45, pp. 127–141). Chicago Linguistic Society.
Harris, J., Neasom, N., & Tang, K. (2016). *Phonotactics with [awt] rules: the learnability of a simple, unnatural pattern in English.* (*24th Manchester Phonology Meeting*, University of Manchester, UK.)
Hatiboğlu, V. (1973). *Pekiştirme ve kuralları.* Türk Dil Kurumu Tanıtım Yayınları.
Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, *44*(1), 45–75. doi: 10.1162/LING_a_00119
Hepworth, G., Gordon, I. R., & McCullough, M. J. (2007). Accounting for dependence in similarity data from DNA fingerprinting. *Statistical Applications in Genetics and Molecular Biology*, *6*(1). doi: 10.2202/1544-6115.1212
Inkelas, S., Küntay, A., Sprouse, R., & Orgun, O. (2000). Turkish Electronic Living Lexicon (TELL). *Turkic Languages*, *4*, 253–275.
Kaufman, B. D. (2014). *Learning an unproductive process: Turkish emphatic reduplication* (Unpublished master's thesis). University of California Santa Cruz.
Kelepir, M. (1999). *Emphatic non-identical reduplication in Turkish.* (Talk given at the LSA Annual Meeting)
Kelepir, M. (2000). To be or not to be faithful. In A. Göksel & C. Kerslake (Eds.), *Studies on Turkish and Turkic languages: Proceedings of the ninth international conference on Turkish linguistics* (pp. 11–18). Harrassowitz Verlag.
Kenstowicz, M., & Kisseberth, C. (2014). *Generative phonology: Description and theory*. Academic Press.
Keuleers, E., Brysbaert, M., & New, B. (2010, August). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650. doi: 10.3758/BRM.42.3.643
Keyser, S. J., & Stevens, K. N. (2006). Enhancement and overlap in the speech chain. *Language*, *82*(1), 33–63.
Kılıç, O., & Bozşahin, C. (2013). Selection of linker type in emphatic reduplication: Speaker's intuition meets corpus statistics. In M. Knauff, N. Sebanz, M. Pauen, & I. Wachsmuth (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, pp. 2722–2727). Cognitive Science Society.
Köylü, Y. (2020). Abstract knowledge of emphatic reduplication in Turkish. In *Talk given at the 5th workshop on Turkic and languages in contact with Turkic (Tu+ 5), University of Delaware.*
Lewis, G. L. (1967). *Turkish grammar*. Oxford University Press.
McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, *17*(2), 207–263.
McCarthy, J. J., & Prince, A. (1993). Generalized alignment. In G. Booij & J. Van Marle (Eds.), *Yearbook of morphology 1993* (pp. 79–153). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-017-3712-8_4
McCarthy, J. J., & Prince, A. (1994). The emergence of the unmarked: Optimality in prosodic morphology. In M. Gonzàlez (Ed.), *Proceedings of the North East Linguistics Society* (Vol. 24, pp. 333–379). Graduate Linguistics Students Association, University of Massachusetts Amherst.
McMullin, K., & Burness, P. (2021). Tier-based modeling of gradience and distance-based decay in phonological processes. In H. Björklund & F. Drewes (Eds.), *Proceedings of the*

17th meeting on the mathematics of language (pp. 50–63). Association for Computational Linguistics.
Müller, H.-G. (2003). *Morphophonologische untersuchungen an reduplikationen im Türkischen* (Unpublished doctoral dissertation). Universität Tübingen.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57. doi: 10.1037/0096-3445.115.1.39
Pater, J. (2016). Universal grammar with weighted constraints. In J. McCarthy & J. Pater (Eds.), *Harmonic grammar and harmonic serialism* (pp. 1–46). London: Equinox.
Pierrehumbert, J. (1993). Dissimilarity in the Arabic verbal roots. In A. J. Schafer (Ed.), *Proceedings of the North East Linguistics Society* (Vol. 23, pp. 367–381). Graduate Linguistics Students Association, University of Massachusetts Amherst.
R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`
Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*(2), 123-138. doi: 10.1037/0096-3445.125.2.123
Rose, S., & Walker, R. (2004). A typology of consonant agreement as correspondence. *Language*, 475–531. doi: 10.1353/lan.2004.0144
Skousen, R., Lonsdale, D., & Parkinson, D. B. (2002). *Analogical modeling: An exemplar-based approach to language* (Vol. 10). John Benjamins Publishing.
Sofu, H. (2005). Acquisition of reduplication in Turkish. In B. Hurch & V. Mattes (Eds.), *Studies on reduplication* (pp. 493–509). Berlin & New York: Mouton de Gruyter.
Sofu, H., & Altan, A. (2008). Partial reduplication: revisited. In S. Ay, Özgür Aydin, İclâl Ergenç, S. Gökmen, S. İşsever, & D. Peçenek (Eds.), *Essays on Turkish Linguistics: Proceedings of the 14th International Conference on Turkish Linguistics* (pp. 63–73). Harrassowitz Verlag.
Stachowski, K. (2014). *Standard turkic C-type reduplications*. Jagiellonian University Press.
Stanton, J. (2017). Segmental blocking in dissimilation: An argument for co-occurrence constraints. In K. Jesney, C. O'Hara, C. Smith, & R. Walker (Eds.), *Proceedings of the 2016 Annual Meetings on Phonology* (Vol. 4). Linguistic Society of America.
Stevens, K. N., & Keyser, S. J. (1989). Primary features and their enhancement in consonants. *Language*, 81–106. doi: 10.2307/414843
Suzuki, K. (1998). *A typological investigation of dissimilation* (Unpublished doctoral dissertation). The University of Arizona.
Taneri, M. (1990). A Type of Reduplication in Turkish. In I. Lee & S. Schiefelbein (Eds.), *Kansas Working Papers in Linguistics* (Vol. 15, pp. 93–126). Linguistics Graduate Student Association, University of Kansas.
Tang, K. (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics*, *24*, 208–214.
Tang, K., & Baer-Henney, D. (2023). Modelling L1 and the artificial language during artificial language learning. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *14*(1), 1–54. doi: 10.16995/labphon.6460
Tang, K., & de Chene, B. (2014). *A new corpus of colloquial Korean and its applications.* (*The 14th Conference on Laboratory Phonology*, Tachikawa, Tokyo, Japan.)
Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249-267. doi: 10.1016/j.wocn.2018.09.004
Wedel, A. (1999). Turkish emphatic reduplication. *Phonology at Santa Cruz (PASC)*, *6*.
Wedel, A. (2000). Perceptual distinctiveness in Turkish emphatic reduplication. In

R. Billerey-Mosier & B. Lillehaugen (Eds.), *WCCFL 19: Proceedings of the 19th West Coast Conference on Formal Linguistics* (Vol. 19, pp. 546–559). Somerville, MA: Cascadilla Press.

Winter, B. (2019). *Statistics for linguists: An introduction using R.* Routledge.

Wissmann, M., Toutenburg, H., & Shalabh. (2007). *Role of categorical variables in multicollinearity in the linear regression* (Tech. Rep. No. 008). Munich, Germany: Department of Statistics, University of Munich. Retrieved from `https://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf`

Yavaş, M. (1980). *Borrowing and its implications for Turkish phonology* (Unpublished doctoral dissertation). University of Kansas.

Yip, M. (1997). Repetition and its avoidance: The case of Javanese. In K. Suzuki & D. Elzinga (Eds.), *Proceedings of the 1995 Southwestern workshop on Optimality Theory (SWOT)* (Vol. 5, pp. 238–262).

Yu, A. (1998). *Prespecification and dissimilation in Optimality Theory: The Case of Turkish Emphatic Reduplication* (Bachelor's Thesis). University of California, Berkeley.

Yu, A. (1999). *Dissimilation and allomorphy: The case of Turkish emphatic reduplication.* University of California, Berkeley Ms.

Zymet, J. (2014). Distance-based decay in long-distance phonological processes. In U. Steindl et al. (Eds.), *Proceedings of the 32nd west coast conference on formal linguistics* (pp. 72–81). Cascadilla Proceedings Project.

Zymet, J. (2018). *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning* (Unpublished doctoral dissertation). University of California, Los Angeles.

Zymet, J. (2019). Learning a frequency-matching grammar together with lexical idiosyncrasy: Maxent versus hierarchical regression. In K. Hout, A. Mai, A. McCollum, S. Rose, & M. Zaslansky (Eds.), *Proceedings of the annual meetings on phonology* (Vol. 7). Linguistic Society of America.

6. Appendix

A. Feature chart

Table 7. Phonological feature values of Turkish consonants (Erguvanlı Taylan, 2015). Abbreviations: son: sonorant, cont: continuant, strid: strident, ant: anterior, cor: coronal, lab: labial, lat: lateral, and nas: nasal

	p	b	t	d	k	g	tʃ	dʒ	f	v	s	z	ʃ	ʒ	ɣ	h	m	n	l	r	j
son	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	+	+	+
voice	–	+	–	+	–	+	–	+	–	+	–	+	–	+	+	–	+	+	+	+	+
cont	–	–	–	–	–	–	–	–	+	+	+	+	+	+	+	+	–	–	+	+	+
strid	–	–	–	–	–	–	+	+	+	+	+	+	+	+	–	–	–	–	–	–	–
ant	+	+	+	+	–	–	–	–	+	+	+	+	–	–	–	–	+	+	–	+	–
cor	–	–	+	+	–	–	+	+	–	–	+	+	+	+	–	–	–	+	–	+	–
lab	+	+	–	–	–	–	–	–	+	+	–	–	–	–	–	–	+	–	–	–	–
high	–	–	–	–	+	+	+	+	–	–	–	–	+	+	+	–	–	–	+	–	+
back	–	–	–	–	+	+	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–
lat	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–
nas	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	+	–	–	–

B. Variable exclusion

Aestheticising: Demircan (1987) proposed a factor called distributional balance or aestheticising. This factor states that speakers would choose a linking consonant that would balance the distribution of features or segments across the reduplicated form. This factor is related to the total identity and partial identity variables. An avoidance in identical features or segments to those of the base form will maintain a balance distribtuion of segments or features. In other words, given the number of all possible features/segments in the reduplicated form, the best linking consonant would minimise an excessive presence of any feature/segment. Under a different interpretation, Wedel (1999) suggests that there is a tendency to balance the word-level sonority, such that speakers would select a linking consonant to balance the overall sonority of the reduplicated form. In the current study we excluded this factor since our focus was on the identity avoidance effect and it has been partially captured by our existing identity variables.
Avoiding phrase formation: Demircan (1987) observed that speakers would avoid a reduplicated form that forms a meaningful phrase. For instance, Demircan suggests that the reason *tatlı* 'sweet' would be reduplicated as *ta**p**-tatlı* and not *ta**m**-tatlı* is because the latter would result in a reduplicated form, i.e., *tam* 'full, exact', which is an existing lexical item in Turkish. For this reason, *-m* is avoided as the LC in this particular instance. In the current study we excluded this factor since our focus was on the identity avoidance effect and this was left for future work.
Vowel height effect: Wedel (1999) proposed an articulatory markedness factor concerning the height of the first vowel in the base form and the linking consonant. The observation was that [s] and [m] are used when the first vowel is a high/mid vowel, while only [s] is used in the context of a low vowel. The articulatory explanation is that the lip closure required for [m] is further away from the jaw position required for a low vowel than for a high or mid vowel. While we have considered articulatory/perceptual markedness, our inclusion of transitional phonotactic probability covers only the linking consonant and the initial consonant. Following almost all previous studies (e.g., Yu 1999), vowel features were excluded in the current study and were left for future work.

C. Descriptive statistics

Table 8. Mean (μ) and standard deviation (σ) of acceptability ratings by item groups (rows) and by linking consonants (columns)

Item group	[p]		[s]		[m]		[r]	
	μ	σ	μ	σ	μ	σ	μ	σ
C_1C_2	4.8238	2.5023	3.4090	2.4937	3.3466	2.4466	1.5124	1.2405
$C_1C_2C_3$	4.8386	2.4124	3.5218	2.5124	3.1855	2.3566	1.8382	1.7104
$C_1C_2C_3C_4$	4.7639	2.4231	4.0388	2.4842	2.8835	2.1946	2.0113	1.8725

Table 9. Descriptive statistics of variables for each of the three item groups. T: true; F: false; μ: mean; σ: standard deviation

	Variable	C_1C_2	$C_1C_2C_3$	$C_1C_2C_3C_4$
	Response	μ: 3.273, σ: 2.526	μ: 3.346, σ: 2.508	μ: 3.424, σ: 2.491
C_1	Total identity	T: 398, F: 6526	T: 541, F: 8903	T: 216, F: 4732
	Sum featural identity	μ: 1.423, σ: 1.235	μ: 1.488, σ: 1.139	μ: 1.362, σ: 1.098
	Sonorant identity	T: 318, F: 6606	T: 576, F: 8868	T: 418, F: 4530
	Voice identity	T: 1636, F: 5288	T: 2338, F: 7106	T: 1470, F: 3478
	Continuant identity	T: 796, F: 6128	T: 1672, F: 7772	T: 674, F: 4274
	Strident identity	T: 357, F: 6567	T: 910, F: 8534	T: 247, F: 4701
	Anterior identity	T: 4392, F: 2532	T: 5156, F: 4288	T: 2460, F: 2488
	Coronal identity	T: 1542, F: 5382	T: 2554, F: 6890	T: 974, F: 3974
	Labial identity	T: 732, F: 6192	T: 842, F: 8602	T: 494, F: 4454
	Nasal identity	T: 79, F: 6845	–	–
C_2	Total identity	T: 563, F: 6361	T: 631, F: 8813	T: 572, F: 4376
	Sum featural identity	μ: 1.792, σ: 1.479	μ: 2.046, σ: 1.342	μ: 2.325, σ: 1.474
	Sonorant identity	T: 1384, F: 5540	T: 2440, F: 7004	T: 1402, F: 3546
	Voice identity	T: 2284, F: 4640	T: 3842, F: 5602	T: 2154, F: 2794
	Continuant identity	T: 2192, F: 4732	T: 3072, F: 6372	T: 1792, F: 3156
	Strident identity	T: 452, F: 6472	T: 616, F: 8828	T: 288, F: 4660
	Anterior identity	T: 3736, F: 3188	T: 5792, F: 3652	T: 3760, F: 1188
	Coronal identity	T: 2044, F: 4880	T: 2090, F: 7354	T: 1558, F: 3390
	Labial identity	T: 238, F: 6686	T: 1142, F: 8302	T: 416, F: 4532
	Nasal identity	T: 79, F: 6845	T: 329, F: 9115	T: 134, F: 4814
C_3	Total identity	–	T: 285, F: 9159	T: 206, F: 4742
	Sum featural identity	–	μ: 1.444, σ: 1.470	μ: 1.668, σ: 1.420
	Sonorant identity	–	T: 2064, F: 7380	T: 1388, F: 3560
	Voice identity	–	T: 2572, F: 6872	T: 1880, F: 3068
	Continuant identity	–	T: 2216, F: 7228	T: 1242, F: 3706
	Strident identity	–	T: 488, F: 8956	T: 175, F: 4773
	Anterior identity	–	T: 3652, F: 5792	T: 1980, F: 2968
	Coronal identity	–	T: 1990, F: 7454	T: 1090, F: 3858
	Labial identity	–	T: 320, F: 9124	T: 250, F: 4698
	Nasal identity	–	T: 339, F: 9105	T: 248, F: 4700
C_4	Total identity	–	–	T: 166, F: 4782
	Sum featural identity	–	–	μ: 1.483, σ: 1.666
	Sonorant identity	–	–	T: 1002, F: 3946
	Voice identity	–	–	T: 1334, F: 3614
	Continuant identity	–	–	T: 410, F: 4538
	Strident identity	–	–	T: 206, F: 4742
	Anterior identity	–	–	T: 2508, F: 2440
	Coronal identity	–	–	T: 1160, F: 3788
	Labial identity	–	–	T: 254, F: 4694
	Nasal identity	–	–	T: 462, F: 4486
	Transitional phonotactic probability	μ: 4.737, σ: 1.201	μ: 4.576, σ: 1.177	μ: 4.506, σ: 1.155

D. Pairwise association

Table 10 summarises the pairwise association results between the response variable and each of the predictors in Section 2.3.1 for each of the three item groups. In the two-consonant group, the total identity variables and sum featural identity variables of C_1 and C_2 are all significant and have negative coefficients. This suggests that they have an identity avoidance effect on the linking consonants as expected. Most individual featural identity variables are significant, except for the anterior identity of C_1, the strident identity, the anterior identity and the coronal identity of C_2. Of the significant individual featural identity variables, only the voice identity of C_1 ($\beta = 0.4045$) has a positive coefficient.

In the three-consonant group, the total identity variables and sum featural identity variables of C_1, C_2 and C_3 are all significant and have negative coefficients. All individual featural identity variables are significant, except for the anterior identity of C_2, the continuant identity, the anterior identity and the coronal identity of C_3. All of the significant individual featural identity variables have negative coefficients.

In the four-consonant group, the total identity variables and sum featural identity variables are all significant, except for the total identity of C_4. Amongst the significant total identity variables and sum featural identity variables, only the sum featural identity of C_3 ($\beta = 0.1581$) has a positive coefficient. All but six individual featural identity variables were insignificant – the nasal identity of C_2, the sonorant identity and the anterior identity of C_3, and the voice identity, the anterior identity and the labial identity of C_4. Of the significant individual featural identity variables, six have positive coefficients – the strident identity ($\beta = 0.4654$) and the anterior identity of C_2 ($\beta = 0.4449$), the continuant identity ($\beta = 01.2100$), the strident identity ($\beta = 0.9814$), and the coronal identity of C_3 ($\beta = 0.6786$) and the sonorant identity of C_4 ($\beta = 0.2887$).

The examination of these pairwise associations indicates that a vast majority of the identity variables (total, sum or individual) have the expected identity avoidance effect. The unexpected effect of identity preference (the opposite of identity avoidance) is most systematic in C_3 in the four-consonant group since it not only has the highest number of variables with a positive coefficient, but also their joined effect was enough to drive the direction of the sum featural identity variable ($\beta = 0.1581$).

Table 10. Pairwise association between the response variable and the predictors for each of the three item groups. β: coefficient; ΔAIC: AIC_{subset} - $\mathrm{AIC}_{superset}$; statistical significance is denoted by $*$ if $\Delta\mathrm{AIC} > 2$; significant positive coefficients of identity variables are in bold.

		C_1C_2		$C_1C_2C_3$		$C_1C_2C_3C_4$	
	Variable	β	ΔAIC	β	ΔAIC	β	ΔAIC
C_1	Total identity	-2.7295	484.48*	-3.3533	887.93*	-3.5667	455.92*
	Sum featural identity	-0.7498	328.65*	-0.8825	694.31*	-0.8825	344.53*
	Sonorant identity	-0.8756	27.25*	-0.3972	8.46*	-0.9690	37.38*
	Voice identity	**0.4045**	16.18*	-0.7797	82.03*	-0.6439	27.57*
	Continuant identity	-1.8726	264.66*	-1.8134	404.03*	-1.0103	60.91*
	Strident identity	-2.8421	406.08*	-2.3863	533.07*	-1.8935	121.56*
	Anterior identity	-0.0581	-1.80	-0.2134	4.6793*	-0.1465	-1.1464
	Coronal identity	-1.3305	171.48*	-0.7432	75.96*	-1.3190	119.26*
	Labial identity	-1.6677	202.46*	-2.2203	415.52*	-2.6013	314.25*
	Nasal identity	-2.2434	65.15*	–	–	–	–
C_2	Total identity	-0.8473	58.78*	-0.9739	93.90*	-0.2840	3.46*
	Sum featural identity	-0.9713	338.90*	-0.9515	577.13*	-0.4508	45.29*
	Sonorant identity	-1.6623	259.66*	-0.3711	19.25*	-0.5162	17.98*
	Voice identity	-1.7804	289.21*	-1.0600	94.13*	-1.9901	125.83*
	Continuant identity	-0.8028	62.24*	-1.0975	160.63*	-0.4688	11.29*
	Strident identity	-0.1831	-0.05	-1.5609	184.68*	**0.4654**	6.18*
	Anterior identity	0.0626	-1.75	0.0160	-1.97	**0.4449**	3.62*
	Coronal identity	-0.1313	-0.14	-1.3716	252.17*	-0.2486	2.36*
	Labial identity	-3.4130	304.18*	-1.8492	318.23*	-1.0904	44.91*
	Nasal identity	-1.8311	42.45*	-0.9771	42.34*	0.2067	-1.12
C_3	Total identity	–	–	-0.9235	35.30*	-0.6774	13.14*
	Sum featural identity	–	–	-0.3898	77.65*	**0.1581**	7.13*
	Sonorant identity	–	–	-0.8435	92.46*	0.0436	-1.86
	Voice identity	–	–	-0.5918	46.94*	-0.6155	19.26*
	Continuant identity	–	–	0.1541	1.63	**1.2100**	107.80*
	Strident identity	–	–	-0.3390	5.64*	**0.9814**	22.67*
	Anterior identity	–	–	0.0979	0.70	-0.1583	-1.24
	Coronal identity	–	–	-0.1611	1.84	**0.6786**	33.35*
	Labial identity	–	–	-1.5442	90.21*	-0.7171	11.98*
	Nasal identity	–	–	-1.1722	65.84*	-0.8816	24.12*
C_4	Total identity	–	–	–	–	-0.1119	-1.67
	Sum featural identity	–	–	–	–	-0.3054	31.43*
	Sonorant identity	–	–	–	–	**0.2887**	4.03*
	Voice identity	–	–	–	–	-0.0537	-1.78
	Continuant identity	–	–	–	–	-1.1192	52.52*
	Strident identity	–	–	–	–	-1.6822	81.85*
	Anterior identity	–	–	–	–	-0.2237	-0.02
	Coronal identity	–	–	–	–	-0.8604	54.17*
	Labial identity	–	–	–	–	-0.2263	-0.57
	Nasal identity	–	–	–	–	-0.3971	5.83*
	Transitional phonotactic probability	0.2174	28.54*	-0.0110	-1.88	0.0292	-1.62

E. Random effects summaries

Table 11. Random effects summary for Study I (two-consonant base forms).

	Standard Deviation
Participant (Intercept)	0.5994
Base form (Intercept)	0.8775
Word shape (Intercept)	0.9995
Linking consonant (Intercept)	0.2457

Table 12. Random effects summary for Study I (three-consonant base forms).

	Standard Deviation
Participant (Intercept)	0.5859
Base form (Intercept)	0.7890
Word shape (Intercept)	0
Linking consonant (Intercept)	0.8209

Table 13. Random effects summary for Study I (four-consonant base forms).

	Standard Deviation
Participant (Intercept)	0.6606
Base form (Intercept)	1.1707
Word shape (Intercept)	0.2133
Linking consonant (Intercept)	1.5917

F. Model comparison for feature specificity

Table 14. Model comparison for feature specificity. ΔAIC: AIC_{subset} - $AIC_{superset}$

Drop from total and individual	C_1C_2	$C_1C_2C_3$	$C_1C_2C_3C_4$
Total identity	87.07	291.94	249.08
Individual featural identity	1016.18	1791.52	1094.96
Drop from total and sum	**C_1C_2**	**$C_1C_2C_3$**	**$C_1C_2C_3C_4$**
Total identity	150.88	376.39	339.42
Sum featural identity	461.79	1028.13	473.5

Table 15. Model comparison for feature specificity: AIC

	C_1C_2	$C_1C_2C_3$	$C_1C_2C_3C_4$
Total and individual featural identity	**28853.63**	**39131.32**	**20298.78**
Total and Sum featural identity	29408.02	39894.71	20920.24
Individual featural identity	28940.70	39423.26	20547.86
Sum featural identity	29558.90	40271.10	21259.66
Total identity	29869.81	40922.84	21393.74

G. Preference hierarchy

Demircan (1987) observed that the majority of the consonant-initial forms reduplicate with [p]. On the other hand, [m] and [s] are similarly frequent, around half of [p]. In Demircan's list, [p] was preferred for 48, whereas [s] for 29, and [m] for 24 items. There are near-minimal pairs such as [basbajat] 'very stale' and [bembejaz] 'very white' that suggest that the two LCs are not exclusive of each other. [r] was the rarest LC, with only 6 items in Demircan's list.

Yu (1999) compiled an extended corpus of 152 attested emphatic adjectives. 123 forms were taken from Hatiboğlu (1973). 121 of the 152 items were consonant initial. It was found that 46% of the forms reduplicate with [p], 29% with [s], 18% with [m], and 7% with [r]. This distribution supports the preference hierarchy. Similarly, Kelepir (2000); Taneri (1990) reached the same conclusion regarding the preference hierarchy [p] > [s] > [m] > [r].

Given that the previous observations were almost exclusively based on forced choice responses (but see Demir (2018) for open-set response task, and Yu (1998) for a small acceptability judgement task), the preference hierarchy was examined using our acceptability ratings from a larger set of items. Our data confirmed the preference hierarchy from Yu (1999). Table 8 shows that all three item groups have the same preference hierarchy – [p] > [s] > [m] > [r]; however, the mean difference between [s] (3.4090) and [m] (3.3466) is small for the two-consonant group.

To test if the preference hierarchy still holds after factoring in the fixed and random effects, we examined also the random intercepts of the linking consonants of the three models (Tables 11, 12 and 13). The random intercepts of the two-consonant model suggest the same ranking as the raw ratings – [p] (0.0054) > [s] (-0.1854) > [m] (-0.5839) > [r] (-2.1823). However, the random intercepts of the three-consonant model shows a different ranking – [s] (-1.9795)> [m] (-2.2302) > [p] (-2.4501) > [r] (-3.8005). The random intercepts of the four-consonant model again shows a different ranking – [s] (-4.3475)> [p] (-4.5443) > [m] (-5.0474) > [r] (-7.7329). While these rankings do not entirely match the preference hierarchy based on raw ratings, there remains a few similarities: In all three item groups, [p] is preferred over [r], [s] is preferred over [m] and [r] is the least preferred linking consonant. This dispreference for [r] even when we factored in the identity variables supports Wedel (1999, 2000)'s conclusion that the reduplicant with [r] might be lexicalised.

H. By-item acceptability rating

To assess the inter-rater reliability of the ratings, we computed the split-half reliability estimates. The split-half reliability is the split-half correlation, corrected with the Spearman-Brown formula. For each group of participants who completed the same list of words, participants were randomly divided into two equally-sized subgroups. To obtain a stable estimate of the split-half reliability, the *splithalf.r* function from the *multicon* library was used to compute the mean split-half reliability with 1,000 random splits. Overall, the inter-rater reliability were high with the mean reliability of .984 averaged across the five lists (ranging from .980 to .988). This indicates that there is a high degree of agreement among the participants and that our ratings are reliable.

Ratings were first standardised for each participant to remove by-participant variations. Means and standard deviations of the standardised ratings were then computed for the four linking consonants for each word. The dominant linking consonant (LC) is shown in the second column, following by the means and standard deviations of the standardised ratings for [p], [m], [s] and [r]. The words in bold (24 words) were expected to be more variable concerning the expected linking consonants based on a meta-analysis of previous studies (not reported here) as well as the linguistic intuition of one of the authors who is a native Turkish speaker. Contrary to expectations, only a minority (six) of these 24 words were particularly variable with two linking consonants with a mean difference in standardised acceptability rating of smaller than 0.5.

		Mean				Standard deviation			
Word	LC	[p]	[m]	[s]	[r]	[p]	[m]	[s]	[r]
başka	m	-0.5272	1.6196	-0.4173	-0.8359	0.6771	0.3509	0.5384	0.2281
bayağı	s	-0.6010	-0.6745	1.2934	-0.8319	0.6144	0.5347	0.5357	0.4621
bayat	s	-0.4654	0.1436	1.0538	-0.8986	0.7579	0.8043	0.6283	0.3457
bedava	s	-0.6607	-0.3800	0.5600	-0.9489	0.5248	0.7004	0.7927	0.2416
bej	m	-0.2402	0.4016	-0.0804	-0.8608	0.8686	0.9356	0.8742	0.3428
belli	s	-0.6313	-0.5485	1.5664	-0.7319	0.4811	0.3847	0.3420	0.2823
beraber	s	-0.4466	-0.5378	0.4468	-0.6615	0.6261	0.5198	0.8887	0.4610
berrak	s	-0.3011	0.1880	1.0162	-0.6831	0.6475	0.8784	0.7776	0.4612
beter	s	-0.6782	-0.0740	1.1602	-0.6813	0.5944	0.9464	0.7204	0.5656
beyaz	m	-0.5071	1.3456	0.0340	-0.8907	0.5331	0.4213	0.6924	0.2569
bok	m	-0.7463	1.2729	0.0025	-0.8361	0.5028	0.4598	0.7243	0.4183
bol	s	-0.7065	0.0229	0.9685	-0.8326	0.5494	0.8535	0.6131	0.3949
boş	m	-0.2967	1.4897	-0.3665	-0.8080	0.8051	0.4472	0.7493	0.3446
boz	m	-0.1174	0.3516	-0.4442	-0.7480	0.8554	1.0456	0.5907	0.5002
bulanık	s	-0.4127	-0.3009	0.9995	-0.8000	0.7906	0.7058	0.6901	0.4871
buruşuk	s	-0.1380	0.6910	0.7139	-0.7562	0.8461	0.9423	0.7713	0.4811
bütün	s	-0.5542	-0.3716	1.5462	-0.7274	0.4902	0.6450	0.3382	0.3904
büyük	s	-0.4886	0.0006	1.2179	-0.9242	0.5782	0.7911	0.5945	0.2743
çabuk	r	0.0352	-0.7324	0.0961	1.1485	0.8926	0.3973	0.7523	0.6907
canlı	p	1.4464	-0.5808	0.1484	-0.8038	0.2958	0.4721	0.8883	0.3856
cavlak	s	0.3044	-0.5002	0.7277	-0.6720	0.8434	0.6169	0.8701	0.5505
çevik	p	0.6888	-0.5385	-0.0884	-0.0955	0.8031	0.6598	0.8318	0.8329
çevre	p	1.0130	-0.4751	-0.0254	-0.6744	0.6906	0.6471	0.7956	0.6052
çiğ	p	1.0651	-0.1656	-0.3477	-0.6815	0.6465	0.8452	0.7296	0.4188
çirkin	p	1.2560	0.0527	-0.2878	-0.8237	0.5273	0.8133	0.8329	0.3506
cıbıl	s	0.4197	-0.4116	1.0832	-0.4875	0.8917	0.7060	0.7428	0.5949
cılız	p	1.2310	0.0086	0.2155	-0.6714	0.4854	0.8190	0.8491	0.4142
cılk	p	0.3112	-0.4685	-0.1288	-0.6153	0.8095	0.6463	0.8739	0.5271
çıplak	r	-0.2002	-0.3860	0.3361	1.1594	0.9205	0.6213	0.8445	0.7756
cıvık	p	0.8665	-0.0122	0.6507	-0.4215	0.8172	0.8147	0.9209	0.7738
çürük	p	1.2022	0.0527	0.1902	-0.8069	0.5198	0.8242	0.8003	0.4266
dağınık	p	1.3091	-0.5433	0.3923	-0.7048	0.4777	0.5702	0.8929	0.4841
dar	p	1.5300	-0.1174	0.4913	-0.7610	0.3165	0.8252	0.9321	0.4001

dazlak	p	0.5757	0.3225	-0.3013	-0.6286	1.0418	0.9839	0.7352	0.5018
derin	p	1.2945	-0.5726	-0.0784	-0.8123	0.4666	0.5872	0.7220	0.4719
dik	m	0.3626	1.3428	-0.3305	-0.7847	0.8213	0.4124	0.7288	0.4241
dinç	p	0.9836	0.3556	-0.3136	-0.8170	0.8117	0.9059	0.7108	0.3960
diri	p	1.4334	-0.1021	-0.0163	-0.8585	0.3165	0.7112	0.7368	0.4235
dızlak	m	0.6966	1.3603	-0.2669	-0.5695	0.7603	0.4734	0.7890	0.5697
doğru	s	1.0487	-0.5547	1.4758	-0.7666	0.6690	0.5418	0.5317	0.3085
dolu	p	1.4137	-0.5768	0.5571	-0.9340	0.4599	0.5489	0.8254	0.2045
durgun	p	1.3010	-0.2371	0.2990	-0.8838	0.4297	0.6929	0.7683	0.2943
duru	p	1.3569	-0.3786	0.2959	-0.8380	0.4222	0.4861	0.7237	0.3416
düz	m	0.7569	1.5026	-0.6168	-0.8014	0.8346	0.3918	0.4863	0.2233
düzgün	p	1.3146	0.7020	-0.3383	-0.8274	0.4609	0.7643	0.6838	0.4272
gece	p	0.5592	-0.6303	-0.2122	-0.7771	0.9789	0.4135	0.8081	0.2988
genç	p	1.2771	-0.3042	-0.3205	-0.9137	0.5022	0.6869	0.7372	0.3266
geniş	p	1.3126	0.0137	0.1912	-0.6325	0.4397	0.8243	0.9420	0.6277
gergin	p	1.2793	-0.5253	0.4275	-0.7245	0.5818	0.6338	0.9363	0.3765
gevşek	p	1.0648	-0.3506	0.4326	-0.7233	0.6232	0.7677	0.8272	0.5028
gök	p	0.1522	-0.1676	-0.1763	-0.7611	1.0003	0.9705	0.8941	0.6020
güdük	p	0.7903	-0.1806	0.2834	-0.8007	0.7356	0.7660	0.9418	0.4695
gündüz	p	0.8907	-0.3302	-0.1542	-0.7714	0.5884	0.6578	0.7118	0.5359
gür	p	1.0520	0.0114	-0.1255	-0.4805	0.5389	0.7853	0.7678	0.7025
güzel	p	1.1055	-0.4194	-0.0485	-0.8884	0.5622	0.7192	0.7415	0.3807
kalın	p	1.5471	-0.4334	0.7206	-0.7137	0.3194	0.6333	0.8293	0.4002
kara	p	1.3319	-0.4841	-0.1315	-0.8095	0.4611	0.6091	0.7146	0.6020
katı	s	0.7612	-0.4464	1.2112	-0.6288	0.7931	0.5947	0.5680	0.6422
kel	p	0.7799	-0.1881	-0.0547	-0.8561	0.7486	0.7986	0.8262	0.3548
kirli	p	1.3843	-0.5026	-0.0559	-0.9153	0.3125	0.5447	0.7541	0.2866
kırmızı	p	1.4369	-0.2852	-0.3294	-0.8618	0.2934	0.6113	0.6573	0.3834
kısa	p	1.4423	-0.0877	-0.6604	-0.7204	0.5973	0.7751	0.5572	0.3930
kıvrak	s	0.4133	-0.4943	1.4245	-0.8519	0.9090	0.5421	0.5067	0.2927
kızıl	p	1.5234	-0.0881	-0.4858	-0.7440	0.2965	0.7017	0.5689	0.2016
koca	s	0.2832	-0.6989	1.3099	-0.7407	0.8830	0.4729	0.3979	0.5213
kocaman	s	0.4104	-0.6907	1.3737	-0.6524	0.7057	0.3875	0.3114	0.5568
kolay	p	1.0792	-0.5073	0.7770	-0.7694	0.7753	0.5156	0.9664	0.3609
kör	p	0.8624	-0.0364	0.3306	-0.9499	0.6911	0.7667	0.7409	0.3276
kötü	p	0.8013	-0.3623	0.3379	-0.5327	0.8109	0.5784	0.8854	0.5613
kötürüm	s	0.3724	-0.4778	0.4133	-0.3860	0.8482	0.5534	0.8766	0.7614
koyu	p	1.1976	-0.4026	0.5497	-0.8192	0.5864	0.6049	0.7079	0.3706
küçük	p	0.8366	-0.2569	0.4990	-0.8576	0.7812	0.8444	0.8547	0.3512
kuru	p	1.4436	-0.3244	-0.0584	-0.8125	0.2991	0.6007	0.6986	0.4560
mavi	s	-0.3452	-0.6842	1.6016	-0.8121	0.6445	0.4656	0.3007	0.2896
mor	s	-0.7986	-0.9545	1.4349	-0.7711	0.4128	0.3624	0.2955	0.5442
parça	s	-0.7969	-0.0176	0.0679	-0.7554	0.4433	0.8526	0.9114	0.4780
parlak	s	-0.7859	-0.4749	1.3516	-0.6788	0.4842	0.6454	0.3333	0.5260
pembe	s	-0.7016	-0.6653	1.6069	-0.7353	0.2996	0.4203	0.2826	0.2123
perişan	r	-0.6947	-0.3595	0.5774	1.3471	0.5406	0.7047	0.7662	0.3013
pis	m	-0.6734	0.7724	-0.5241	-0.5604	0.6165	0.8927	0.6985	0.5314
sade	p	1.4246	-0.3334	-0.6532	-0.8344	0.5368	0.7837	0.4449	0.2714
sağ	p	0.9654	-0.2231	-0.7780	-0.7672	0.7708	0.8179	0.2907	0.2747
sağlam	p	1.0320	-0.4185	-0.8489	-0.7116	0.6549	0.6581	0.4734	0.5546

salak	p	1.1209	0.0853	-0.9345	-0.6264	0.6390	0.8184	0.2467	0.7473
sarı	p	1.4464	-0.2508	-0.8570	-0.8627	0.2958	0.5570	0.3697	0.3092
sebil	r	0.3619	-0.3128	-0.6422	0.6168	0.8268	0.6449	0.3670	0.9527
sefil	r	-0.0632	-0.3815	-0.9204	1.2958	0.7996	0.6654	0.4321	0.5915
serin	p	1.2388	-0.2668	-0.8685	-0.6031	0.3946	0.6597	0.3173	0.6081
sert	p	1.0825	0.9830	-0.9223	-0.8989	0.7385	0.7518	0.2916	0.4260
silik	p	1.0689	0.2056	-0.9131	-0.9311	0.6468	0.9107	0.3884	0.3320
şirin	p	1.2555	-0.0839	-0.8588	-0.8077	0.4845	0.7613	0.5426	0.3961
sivri	p	1.5424	-0.0103	-0.7782	-0.7256	0.4839	0.8263	0.4123	0.3273
siyah	m	0.3335	1.6189	-0.8115	-0.8330	0.8983	0.3422	0.1677	0.1931
sıcak	m	0.5997	1.1739	-0.8832	-0.8219	0.7545	0.6959	0.4587	0.4464
sığ	p	1.1698	0.2767	-0.7634	-0.7859	0.6081	0.9664	0.2901	0.3149
sık	m	0.1017	0.6797	-0.9448	-0.8777	0.9529	0.8224	0.3276	0.2956
sıkı	m	0.7686	1.3796	-0.8506	-0.6754	0.7837	0.4025	0.3477	0.4949
sıkkın	p	0.5850	-0.0129	-0.9338	-0.9061	0.9756	0.8583	0.2662	0.3387
sıklam	r	-0.2542	-0.3870	-0.7664	0.7790	0.7145	0.6115	0.3380	0.9542
sıska	p	0.8703	0.1227	-0.6780	-0.6583	0.7658	0.8615	0.3467	0.4764
soğuk	p	1.3435	0.2466	-0.8328	-0.7988	0.5328	0.8582	0.2088	0.2108
sulu	p	1.3186	-0.1221	-0.9461	-0.8789	0.3927	0.9096	0.2662	0.3318
takır	m	-0.2366	1.4143	0.0347	-0.7481	0.6656	0.3356	0.7942	0.4582
tam	s	0.0430	-0.5618	0.6996	-0.7265	0.8580	0.6241	0.8670	0.3717
tamam	s	0.3765	-0.3826	1.4137	-0.6222	0.7811	0.7703	0.3252	0.6068
tatlı	p	1.0233	-0.2614	0.0314	-0.7454	0.7978	0.7184	0.7189	0.2723
taze	p	1.3494	-0.0831	-0.0577	-0.8435	0.3750	0.6954	0.7255	0.3690
temiz	r	-0.1526	-0.5765	-0.5371	1.4664	0.7748	0.6744	0.5523	0.3765
tok	p	0.3258	0.2060	0.2081	-0.6886	0.9626	0.9087	0.8946	0.5140
top	s	-0.4289	-0.6914	0.6909	0.3371	0.8491	0.4713	0.8417	1.1187
topaç	s	-0.0306	-0.4422	0.4515	0.1675	0.9311	0.5290	0.8542	0.9788
turuncu	p	1.3806	-0.2706	0.1884	-0.7014	0.5625	0.6650	0.8303	0.2842
tuzlu	p	1.2364	-0.2069	-0.5810	-0.7145	0.6036	0.6001	0.4502	0.4739
yakın	p	1.1253	-0.4153	0.0338	-0.8208	0.6390	0.7589	0.8811	0.3551
yalnız	p	1.3325	-0.3624	-0.4896	-0.7957	0.4028	0.6402	0.5621	0.3361
yanlış	p	1.2304	-0.4160	-0.2698	-0.7577	0.5236	0.5191	0.6731	0.2573
yarık	p	0.4743	-0.3967	0.1801	-0.8025	0.8729	0.7499	0.8266	0.4306
yaş	m	0.4778	0.5367	-0.5728	-0.7907	0.8819	1.0068	0.5589	0.2058
yassı	p	0.8995	0.5195	-0.4644	-0.7472	0.8100	0.9900	0.5962	0.3427
yavaş	p	1.0645	-0.3271	0.2539	-0.9051	0.5633	0.6488	0.8021	0.3266
yeni	p	1.4351	-0.6550	0.0948	-0.9329	0.3034	0.5030	0.7832	0.2950
yeşil	m	0.3812	1.6140	-0.4133	-0.7888	0.9417	0.4783	0.5570	0.1856
yırtık	p	0.9290	-0.3057	0.4909	-0.9081	0.6908	0.7108	0.8667	0.2798
yoğun	p	1.1671	-0.4801	0.4598	-0.7123	0.6378	0.5484	0.9441	0.3963
yorgun	p	1.0091	-0.6846	0.6618	-0.9034	0.6784	0.4155	0.7348	0.2768
yumru	s	0.5817	-0.2130	1.0221	-0.7591	0.7622	0.6680	0.6875	0.4734
yumuşak	s	0.5435	0.3196	0.9072	-0.7306	0.7207	0.8395	0.6894	0.5479
zayıf	p	1.4014	-0.4054	-0.7933	-0.7710	0.3049	0.5109	0.4504	0.5047
zor	p	1.3629	-0.4034	-0.6942	-0.6904	0.5771	0.6219	0.4244	0.4618