Labusch, M., Perea, M., Marcet, A., Baciero, A., & Fernández-López, M. (in press). The Role of Diacritics in the Recognition of Words across Different Writing Systems. In H. Winskel and H. Pae (Eds.*), Handbook of Nonlinear Writing Systems – Complex Processes and Learning Challenges*. Springer

The Role of Diacritics in the Recognition of Words across Different Writing Systems

Melanie Labusch [1,2], Manuel Perea [1,2], Ana Marcet [3], Ana Baciero [4], and Maria Fernández-López [5]

[1] ERI-Lectura and Departamento de Metodología, Universitat de València, Valencia, Spain

[2] Centro de Investigación Nebrija en Cognición, Univesidad Nebrija, Madrid, Spain

[3] Departamento de Didáctica de la Lengua y la Literatura, Universitat de València, Valencia, Spain

[4] Departamento de Psicología Experimental, Procesos Cognitivos y Logopedia, Universidad Complutense de Madrid, Madrid, Spain

[5] Departament de Psicologia Bàsica, Universitat de València, València, Spain

Short title: Diacritics and word recognition

Correspondence:

Manuel Perea

Dpto. Metodología and ERI-Lectura

Av. Blasco Ibáñez, 21

46010-Valencia (Spain)

Email: mperea@uv.es

Abstract

This chapter examines the role of diacritics across diverse writing systems, focusing on their role during lexical access. Through a comparative analysis, we show the varied functions and historical evolution of diacritical marks across languages (e.g., marking different sounds, vowel length, tones, or stress assignment). We also review recent research on the processing of diacritical letters in visual word recognition tasks across various experimental paradigms, which suggests that differences in diacritic function lead to distinct outcomes. Finally, the chapter explores the challenges that diacritical letters pose to computational models of word recognition and examines proposals for representing these letters within these models.

Key words: visual word recognition, reading, diacritics, lexical access, modeling

4.1 Introduction

When we translate sentences like *The children are excited about their summer vacation in the mountains* into other languages using Roman script, we observe the frequent presence of diacritics, each with its own characteristics and serving different linguistic functions. For example: *Los niños están emocionados por sus vacaciones de verano en la montaña* (Spanish)*, Les enfants sont excités par leurs vacances d'été à la montagne* (French)*, A gyerekek izgatottak a nyári vakációjuk miatt a hegyekben* (Hungarian), and *Děti jsou nadšené z prázdnin na horách* (Czech)*.*

Diacritical marks (or simply diacritics) are small additions to letters that are seen occasionally in English (e.g., <café>, <naïve>, <façade>, <resumé>), but they play a significant role in most other writing systems (Comrie, 1996; Wells, 2001). These marks may modify pronunciation, indicate stress patterns, or just differentiate words that otherwise look similar. For instance, without diacritics, distinguishing the Spanish word <sábana> (*sheet*) from <sabana> (*savannah*), or the French word <sûr> (*certain*) from <sur> (*on top*), would be complicated. Indeed, diacritics are key elements of orthography in many languages (see Protopapas & Gerakaki, 2009, for review). Among languages that use the Roman alphabet, English is unusual in its limited and purely optional use of diacritical marks from loanwords, often seen stylistically in brand names (e.g., <Häagen-Dazs>) or musical bands (e.g., <Motörhead>) (see Figure 1).

English (Roman script)
The children are excited about their summer vacation in the mountains.

Serbian (Cyrillic and Roman scripts)
Деца су узбуђена због свог летњег распуста у планинама.
Deca su uzbuđena zbog svog letnjeg raspusta u planinama.

Greek
Τα παιδιά είναι ενθουσιασμένα για τις καλοκαιρινές τους διακοπές στα βουνά.

Arabic
.الأطفال متحمسون لعطلتهم الصيفية في الجبال

Thai
เด็กๆตื่นเต้นกับวันหยุดฤดูร้อนในภูเขา.

Hindi (Devanagari script)
बच्चे पहाड़ों में अपनी गर्मी की छुट्टियों के लिए उत्साहित हैं।

Japanese (Hiragana)
こどもたちはやまのなつやすみにわくわくしています。

Chinese (Simplified and Pinyin scripts)
孩子们对他们在山里的暑假感到兴奋。 **Háizimen duì tāmen zài shānlǐ de shǔjià gǎndào xīngfèn.**

Armenian
Երեխաները ոգևորված են իրենց ամառային արձակուրդներով լեռներում:

Georgian
ბავშვები აღფრთოვანებულნი არიან თავიანთი ზაფხულის არდადეგებით მთებში.

Hangul (Korean)
아이들은 산에서 여름 방학을 기대하고 있어요.

Figure 1. Examples of diacritics across various scripts, with exceptions in English, Chinese, Armenian, Georgian, and Hangul, using the same sentence.

As reviewed by Wells (2001), linguists have often considered diacritics—and, in general, spelling—a minor feature of any language, certainly when set against larger topics in syntax, morphology, or phonological theory. One reason is that diacritics are elements of the surface orthographic form rather than the underlying phonological representations. Consequently, diacritics may be perceived as additional details that do not change anything fundamental in the very essential phonological inventory of a language. Furthermore, in languages where diacritical marks are purely supplementary (e.g., English), their absence may reinforce the belief that they hold a peripheral place. Indeed, in English, context, morphology, and irregular spelling conventions are heavily

relied upon for meaning and pronunciation; thus, this reinforces the Anglocentric argument (Share, 2008) that diacritics are somewhat superfluous.

However, as we intend to show in this chapter, the study of diacritics offers researchers an excellent window into the interface between phonology and orthography, as well as into how letters are represented in the brain. We must keep in mind that diacritics are often explicit markers of pronunciation that help in clarifying how sounds are conventionally represented in writing.

Before we explore the role of diacritics in models of visual word recognition and the relatively limited literature on this topic, it is crucial to first understand the origins and purposes of diacritics. This background may provide insight into the relationship between spoken and written language and also show how writing systems adjust to meet the communication needs of their users. Specifically, we will examine several writing systems in regard to the origins and developments of diacritical marks.

4.1.1 Roman alphabet

The Roman alphabet was initially designed for the Latin language and lacked diacritics. However, it turned out that this alphabet did not have the capacity to represent many sounds from other languages that later adopted this script. To accommodate new phonemes, the different languages using this alphabet modified existing letters using diacritical marks or, in some cases using digraphs (e.g., <sh> for the English phoneme /ʃ/), rather than creating entirely new letters. One of the few exceptions to this rule is the letter <þ> in Icelandic which represents the phoneme /θ/ and originates from the runic alphabet.

The role of diacritics varies greatly across languages using the Roman alphabet. For simplicity's sake, we will cite three examples: Spanish, French, and Hungarian. First, in Spanish, diacritical marks in the form of acute accents signal the stressed syllable of a word (under some norms; e.g., <papa> /'papa/ [*potato, Pope*] versus <papa> /pa'pa [*dad*]) as well as it is used to distinguish the meaning of homographs (e.g., <tú> [*you*, singular] versus <tu< [*your*]). Furthermore, a tilde over the letter <n> (i.e., the letter <ñ>), represents another phoneme, the palatal nasal phoneme /ɲ/, making <mono> (*monkey*) into <moño> (*bun*). Additionally, the diaresis (¨) is used over the letter <u> in <güe> and <güi> to indicate that the letter <u> should be pronounced. For example, in <pingüino> (/piŋˈgwino/ *penguin*), the diaresis signals that the u is pronounced, as opposed to words like <guitarra> (/giˈtara/ *guitar*), where the <u> is silent.

Second, in French, the acute accent (´) marks a closed vowel sound, as seen in <café> (/ka.'fe/ *coffee*), while the grave accent (`) indicates an open vowel sound, as in <père> (/pɛʁ/ *father*). The circumflex (ˆ) can signal a historical change or lengthening of a vowel, as in <forêt> (*forest*). The cedilla <ç> softens a hard <c> into a soft <c> sound before <a>, <o>, or <u>, as in <façon> (/fa.sõ/, *way, manner*) (see Labusch et al., 2023, for a more detailed explanation).

Third, in Hungarian, diacritics are used to represent its complex vowel system and to distinguish between short and long vowels that may determine the meaning of a given word (see Benyhe et al., 2023). Long vowels are indicated by an acute accent (<á, é, í, ó, ú>), while the umlaut (<ö, ü>) and their counterparts with long vowels (<ő, ű>) show front rounded vowels not represented by Latin (e.g., <kor> /kor/*age* versus

<kór> /koːr/ *disease*, differing in vowel length; <sör> /ʃør/ *beer* versus <sor> /ʃor/ *line*, differing in vowel quality).

4.1.2 Cyrillic alphabet

The use of diacritics in the Cyrillic script varies across languages. In Russian, the letter <ë> represents the sound /jo/, unlike <e> /je/. The letter <й> represents the /j/ sound, unlike <и> /i/. Additionally, acute accent marks may, occasionally, indicate lexical stress on vowels in unfamiliar (often loan) words or to clarify homographs (e.g., <за́мок> /ˈzamək/ *castle* and <замо́к> /zɐˈmok/ *lock*). Other languages using Cyrillic, such as Serbian, attach small features to some letters, rather than separate marks to capture specific sounds; for example, <Ћ> /tɕ/ and <Ђ> /dʑ/ from the letter <Т> /t/.

4.1.3 Greek alphabet

Ancient Greek used diacritical marks to indicate pitch and breathing sounds. For instance, accents like the acute (´), grave (`), and circumflex (ˆ) marked tonal variations, while breathing marks (e.g., ʽ) showed aspiration. In Modern Greek, this system was simplified, and currently only the acute accent is used to indicate the stressed syllable in polysyllabic words (e.g., <αγορά> /aɣoˈra/ [*market*]).

4.1.4 Semitic scripts

The origin of Semitic scripts can be traced back to ancient writing systems that developed in the Near East. The most extended Semitic script is the Arabic script, which is used in Arabic and other languages from various families. This script employs dots as diacritical marks to differentiate letters that share the same base shape (see

AlJassmi & Perea, 2024; Boudelaa & Marslen-Wilson, 2010; see also Boudelaa, this Volume). Notably, early Arabic lacked these dots, which led to ambiguity in reading. For instance, the base shape <ب> could be interpreted as different letters depending on the context. Scholars resolved this by adding dots: <ب> (/bā/) with one dot below represents /b/, <ت> (/tā/) with two dots above represents /t/, and <ث> (/thā/) with three dots above represents /θ/. These dots expanded the script's ability to accurately represent phonemes, reducing uncertainty and enhancing readability. Moreover, Arabic words are typically written without short vowels, but there is a system of diacritical marks that can indicate short vowels and other phonetic features. These diacritics, known as *harakat*, are often used in religious texts or educational settings to facilitate the correct pronunciation of words.

4.1.5 Thai script

In Thai, diacritical marks were introduced with the creation of the script in the 14th century and are used for marking the five tones, which in turn may distinguish word meanings. For example, the Thai word <มา> (/maː/, neutral tone) means "to come", while the word <มม้า> (/máː/, high tone) means "horse". Additionally, vowel diacritics may indicate short and long vowels—a similar case occurs in other languages of the same family (e.g., Burmese and Lao).

4.1.6 Devanagari script

Unlike other writing scripts, Devanagari uses additions above, below, or beside the base letters. For instance, in <कि> (/ki/), the diacritic mark for the vowel sound /*i*/ is added to the consonant <क> (/ka/). While, in the previous case, the diacritic changes

the pronunciation of the letters, there are also diacritics for nasalization and for aspiration of vowels. In addition, due to the influence of Persian and Arabic around the 12th and 13th centuries, scholars in India developed diacritical marks to modify consonants of the original Devanagari script for non-native sounds from these languages—they were represented as a small dot beneath the consonant. For instance, the syllable <फ> (/pha/) becomes <फ़> (/fa/), allowing the Devanagari script to represent sounds from Persian and Arabic that are not native to the Sanskrit language (e.g., the sound /f/). This adaptation demonstrates how scripts accommodate linguistic changes over time, reflecting the diverse phonetic landscape of South Asia .

4.1.7 Japanese kana syllabaries

As Chinese characters (*kanji*) were not sufficient for all Japanese sounds, around the 8th and 10th century scholars developed kana syllabaries—*hiragana* and *katakana*—that would represent Japanese phonology in the form of syllables. In the two kana syllabaries, diacritics serve to modify consonant sounds of their syllables without increasing the number of base letters. Specifically, the addition of two small strokes called *dakuten* changes a voiceless consonant to its voiced counterpart (e.g., <か> /ka/) becomes <が> /ga/), while a small circle called *handakuten* indicates a plosive sound (e.g., <は> /ha/ becomes <ぱ> /pa/). Thus, Japanese kana uses diacritics to represent additional sounds in a systematic manner (e.g., changing voiceless to voiced consonant sounds) without the need to create new characters.

4.1.8 Chinese and Pinyin scripts

Logographic Chinese characters do not use diacritics. However, Pinyin, the Romanized system for Mandarin Chinese that is used when learning to read and write Mandarin since the 1950s, employs tone marks to distinguish meanings. The four tones are indicated by diacritical marks: high (¯), rising (ˊ), falling-rising (ˇ), and falling (ˋ). These tone markers are necessary to differentiate words like <mā> (*mother*) from <mǎ> (*horse*), <má> (*hemp*), and <mà> (*scold*), as the same syllable can have entirely different meanings depending on the tone.

4.1.9 The case of Georgian, Armenian, and Hangul scripts

While most languages have incorporated diacritics in one way or another, other languages have developed scripts that do not require them. The alphabets for Armenian (an Indo-European language) and Georgian (a Kartvelian language) are probably the best examples. These two alphabets were independently designed in the Caucasus region around the 5th century, and both scripts have unique letter correspondences for each phoneme.

Hangul is another writing system that does not have diacritics. It is a featural alphabet created in the 15th century to replace the Chinese characters, which were not well-suited to the Korean language. The shapes of the Hangul letters are based on the articulatory features of the sounds they represent (e.g., the consonant shapes reflect the position of the tongue and mouth), making it easy to read and learn Korean . Despite its logical structure, it was not until the 20th century that Hangul was universally adopted across Korea, as Chinese characters remained dominant in official and scholarly writing for centuries.

4.1.10 Brief summary

In this subsection, we have seen that diacritical marks play a crucial role in the orthography of most writing systems. Their evolution reveals the adaptability of writing systems to meet the needs of their speakers. This raises the question: Are these diacritics truly necessary? We will address this topic from an empirical perspective in the next section.

4.2 Empirical findings on the role of diacritics in visual word recognition and reading

Visual word recognition, the process by which readers quickly and efficiently understand the meanings of written words, is a crucial cognitive skill for reading. A key element of this process involves encoding the identity and position of each letter that constitutes a word (for a recent review, see Grainger, 2024). As previously discussed, while diacritics form part of many writing systems, their role can vary considerably not only across but also within languages. In Spanish, for example, the tilde on <ñ> (/ɲ/) represents a distinct phoneme from <n> (/n/), while an accent mark on a vowel (e.g., <á>) does not change the phoneme but indicates the stressed syllable.

In this section, we review empirical findings from various word recognition experiments with adult readers that use different techniques across languages, including masked priming, lexical decision tasks, categorization tasks, and eye movement evidence during reading, all focused on the processing of words with diacritics. For the sake of brevity, we focus on these techniques rather than on reading aloud (for instance, see Protopapas & Geraki, 2009), as the former tasks do not have an explicit phonological component.

4.2.1 Masked Priming: The Role of Diacritics in Early Word Processing

Masked priming (Forster & Davis, 1984) is a paradigm used to explore early processes in word recognition. In this paradigm, a briefly presented (~33ms) forwardly masked prime stimulus, usually in lowercase, precedes the target word, typically in uppercase, allowing researchers to study how such a flashed prime influences the processing of the target (e.g., comparing <house-HOUSE> versus <mouse-HOUSE> to examine early orthographic processing), thereby tapping into the early stages of orthographic and phonological processing (see Grainger et al., 2008). This technique has recently been used in different languages to examine whether letters with diacritical marks have their own representation in the word recognition system or are merely variations of their base letters.

In French, Chetail and Boursain (2019) found that, in a lexical decision task with non-diacritical target words like <TAPER> (*to type*), response times were faster when preceded by identity primes (e.g., <taper>) than by pseudoword primes with an added diacritic (e.g., <tàper>) or by replaced-letter primes (e.g., <tuper>). There was no difference between the latter two conditions, suggesting that diacritical and non-diacritical letters may involve distinct orthographic representations.

Perea et al. (2020b) observed a similar pattern in Spanish for non-diacritical target words. For instance, the identity prime <feliz> resulted in faster response times for the target word <FELIZ> (*happy*) than the pseudoword primes <fáliz> or <féliz>. When diacritical target words (e.g., <FÁCIL> [*easy*]) were used, both identity primes (e.g., <fácil>) and non-diacritical conditions (e.g., <facil>) produced similar response times, both faster than the control priming condition (e.g., <focil>). Marcet et al. (2020) extended these findings to diacritical and non-diacritical consonant target

words. For instance, for the target word <MUÑECA> [*doll*], both the primes <muñeca> and <muneca> were more effective, to a similar degree than the prime <museca> (where <n> was replaced with <s>); in contrast, for the non-diacritical target word <MONEDA> [*coin*] the identity prime <moneda> was more effective than both <moñeda> and <moseda>. The resemblance between the pattern for diacritical vowels (which maintain the same phonemes) and diacritical consonants (which alter their phonemes) in Spanish suggests that the observed effects are due to perceptual elements rather than phonology itself.

In Finnish, Perea et al. (2022a) found that omitting a diacritical mark from a target word (e.g., the prime <poyta> for the target <PÖYTÄ> [table]) did not slow response times compared to the identity prime (<pöytä>), even though the omission caused phoneme changes and vowel disharmony (e.g., the letters <ö > and <a> cannot coexist in the same Finnish monomorphemic word). In addition, the replaced-letter priming condition (e.g., <paytä>) led to slower response times on the target word. Thus, these priming effects appear to be more orthographic than phonological.

Benyhe et al. (2023) conducted a lexical decision experiment in Finnish using the same design as Perea et al. (2020b). Unlike in Spanish, where vowel diacritics indicate the stressed syllable, in Hungarian, vowel diacritics indicate vowel length, which may modify the phoneme (e.g., <mése> [ˈmeːʃɛ] and <mese> [ˈmɛʃɛ]) or not (e.g., <róka> [ˈroːkɒ] and <roka> [ˈrokɒ]). The results mirrored Perea et al.'s (2020b) findings, suggesting again that phonology plays, if anything, a minor role in the priming effects.

Thus, in general, empirical findings across Roman script languages suggest that priming effects are typically not shaped by phonological processes but rather by

perceptual ones. For example, Perea et al. (2023) observed that adding diacritical vowels to English primes (e.g., <nórth> for <NORTH>) incurred processing costs for monolingual U.S. participants, despite their lack of familiarity with diacritical representations.

Regarding other writing systems, in Thai, Winskel and Perea (2014) found that for diacritical target words (e.g., <ห้อง>, *room*), the identity priming condition led to faster lexical decision times than a priming condition with modified tone diacritics (e.g., <ห่อง>). This, in turn, produced faster responses than a priming condition in which the tone was retained, but the base letter was modified (e.g., <ค้อง>). Winskel and Perea (2014) interpreted this pattern as consistent with the idea that tone diacritics are encoded early in Thai and that the identity of the base letter has a stronger role than the tone diacritic alone.

In Arabic, Perea et al. (2016, 2018) found faster response times for target words (e.g., <تفاعل> [interaction]) preceded by identity primes than for pseudoword primes with a different diacritic in one of the letters keeping the basic shape (e.g., <تقاعل>) or with a visually dissimilar letter (e.g. <تخاعل>). Nonetheless, interpreting the null differences between the visually similar and visually dissimilar priming conditions requires caution, as diacritics are the single most important feature in letter discrimination in Arabic (Wiley et al., 2016). Importantly, using masked priming with a letter-level task (the alphabetic decision task), which is generally considered an adaptation of the lexical decision task for the study of letters (see Grainger, 2024), Aljassmi and Perea (2024) found a small advantage in recognizing letters preceded by repeated or diacritical primes over unrelated primes, regardless of whether the letter

contained a diacritic. This suggests that the null effect mentioned above in Arabic is due to higher-level processing rather than letter-level processing.

4.2.2 Single-Presentation Studies

While masked priming experiments explore very early stages of word processing, they do not directly assess the impact of diacritics on word recognition but rather the relationship between diacritical and non-diacritical letters when explicitly presented (see Andrews, 1997; Gómez et al., 2021, for discussion). A more direct method to tackle the study of diacritical words is using unprimed paradigms using lexical decision or semantic categorization tasks.

4.2.2.1 Lexical decision studies

In a go/no-go lexical decision task with Spanish words containing diacritical vowels (e.g., <cárcel> *prison*), Schwab (2015) found similar response times for intact (<cárcel>) and omitted-diacritic (<carcel>) words (see also Marcet et al., 2021, Experiment 2, for a replication with a different set of items). (In a go/no-go task, participants respond to one category [e.g., words] but not to another category [e.g., nonwords].) The experiments were conducted in two blocks: one for accented words/pseudowords and another for unaccented items. Schwab (2015) concluded that accented and non-accented vowels may share orthographic representations in Spanish. Using a standard two-choice lexical decision task with both words and pseudowords, Marcet et al. (2021, Experiment 1) found a word/nonword dissociation: intact words (e.g., <cárcel>) were responded to faster than words with omitted diacritics (e.g., <carcel>), while

accented pseudowords showed the opposite effect, suggesting a bias toward categorizing accented stimuli as words in the standard lexical decision task.

Notably, semantic categorization tasks can provide clearer insights into the role of diacritics in word recognition than the unprimed lexical decision task. The reasons are that the lexical decision task can be performed without unique word identification (Grainger & Jacobs, 1996) and it can be influenced by the visual familiarity of the printed stimulus (see Perea et al., 2020a). Following this logic, several studies have reported that the processing of words with diacritics differs among languages.

4.2.2.2 Semantic Categorization studies

In a semantic categorization task where participants had to decide if a word referred to an animal or not, Perea et al. (2021) found similar response times for intact (e.g., <ratón> *mouse*) and omitted-diacritic words (e.g., <raton>) in Spanish. This pattern replicates, Schwab (2015) and Marcet et al.'s (2021) results, with another paradigm. Relatedly, Labusch et al. (2022) found a minimal cost (about 7 ms) for adding a diacritical mark to a non-diacritical word (e.g., <cebrá> vs. <cebra> *zebra*).

In German, Perea et al. (2022b) found longer response times when diacritics were omitted (e.g., <Kröte> *toad* was processed faster than <Krote>), consistent with the distinct phonemes represented by <ö> and <o> (/ø/ and /o/, respectively), which would entail different orthographic representations (see Ziegler et al., 2000).

In addition, Labusch et al. (2023) found that removing diacritics from French words caused minimal disruption to reading speed, similar to previous findings in Spanish. However, when changing the diacritics of a word from <é> to <è> or vice

versa (<é> and <è> represent close /e/ and open /ɛ/ sounds, the processing cost was larger (e.g., from <chèvre> *goat* to <chévre>). In addition, there was a cost for diacritics that were added to non-diacritical words, particularly when modifying the letter <e> (e.g., <chèval> from the word <cheval> *horse*).

Taken together, these findings indicate that the function of diacritics varies significantly among languages that use the Roman alphabet, depending on how consistently they indicate different phonemes. In languages like Spanish, where diacritics can be omitted without major consequences, their absence generally has little impact on processing speed. In contrast, languages like German, where diacritics represent distinct phonemes, greater processing costs when diacritics are omitted are found. The case of French would be intermediate—in particular, the letters <é> and <è> may behave more like the diacritical letters in German.

4.2.3 Evidence from Eye-Tracking Studies

Eye-tracking technology records an individual's eye movements while reading, providing a more naturalistic way to study reading processes. However, to our knowledge, only one published study has examined the role of diacritics using this method, and only in Spanish. Marcet and Perea (2022) conducted a sentence-reading experiment where target words requiring diacritics (e.g., <cárcel>) were presented either intact or without the diacritical mark (e.g., <carcel>) while maintaining the diacritics in the rest of the sentence . They found minimal disruption in first-fixation and first-pass durations, with only a minor disadvantage for omitted diacritics in total reading time. Based on these findings, Marcet and Perea (2022) suggested that

Spanish readers can easily compensate for missing diacritics using context and internalized phonological rules.

Further research using this paradigm in other languages is needed to better understand the role of diacritics in sentence and text reading.

4.3 Modelling Letters with Diacritics in Visual Word Recognition Across Scripts

Visual word recognition comprises various hierarchical layers of processing, ranging from the visual perception of single-letter features to the activation of whole-word representations (see Dehaene et al., 2005; Grainger et al., 2008). Nearly all computational models of visual word recognition, including the influential Interactive Activation model of McClelland and Rumelhart (1981; Rumelhart & McClelland, 1982), have been developed based on the English alphabet, which lacks diacritical marks. However, as reviewed earlier, many languages use diacritics to specify the pronunciations of letters or to indicate lexical or grammatical contrasts. This poses a challenge for current models, which generally have not accounted for the processing of letters with diacritical marks. In this section, we address the expansion of these models to cover letters with diacritics.

As described earlier, diacritical marks in the Roman script can be used to modify existing letters without changing any phoneme: lexical stress, as in Spanish vowel accents (e.g., <á>); word length, as in Hungarian (e.g., <á>); or tone markers, as in Vietnamese (e.g., <ã>). In principle, these modifications may still allow for shared representations between diacritical and non-diacritical letters. Diacritical marks can also be used to indicate distinct phonemes that require their own letter representations (e.g., <ä> in German).

Interestingly, even within the same language, diacritics can have different functions. For instance, in Spanish, <ñ> represents a distinct phoneme, while <á> only indicates lexical stress. All these differences have to be kept in mind when attempting to model the recognition of written words.

4.3.1 Limitations of Current Models in Handling Diacritics

All traditional leading computational models of visual word recognition, including the interactive activation model and those models that use the same orthographic scheme (e.g. LTRS model: Adelman, 2011; dual-route cascaded model: Coltheart et al., 2001; spatial coding model: Davis, 2010; multiple read-out model: Grainger & Jacobs, 1996; CDP+ model: Perry et al., 2007), have been designed assuming a fixed set of letters-usually the 26 letters of the English alphabet, which were based on the—simplified and unrealistic—14-feature letter system composed of straight lines by Rumelhart and Siple (1974) for uppercase letters. In principle, feature detectors within these models respond to visual features of letters: there are letter detectors that recognize letters across font and case, and word detectors that integrate information across all positions to recognize whole words. However, even within Latin-based alphabets, most languages extend beyond the 26 letters of the English alphabet.

The same problem also occurs in other models that do not use the orthographic scheme of the interactive activation model. For simplicity, the Bayesian reader model (Norris, 1996) assumes that all letters are equally confusable, adopting each a random array of vectors (e.g., <n> and <m> would be equally confusable as <n> and <s>, which goes against current evidence, see Marcet & Perea, 2017, 2018; see also Bae et al.,

2024, for evidence in Korean). The connectionist model proposed by Ans et al. (1998) includes diacritical letters in French at the letter level, but it does not include a feature-letter level—a similar case occurs in the connectionist models proposed by Ziegler et al. (2000) for German (see also Hutzler et al., 2004).

In a recent paper, Snell (2024) indicated that models of visual word recognition have often disregarded the links between the visual features and the letter level, as their focus is more on the interplay between the letter and word levels (see Davis, 2010, Reichle, 2020, for a similar point). While, as Balota et al. (2006) indicated, going from visual features to more abstract representations is an extremely challenging enterprise, it is one that is necessary to fully understand the journey from ink to meaning.

4.3.2 A Proposal for Integrating Diacritics into Models of Word Recognition

As stated above, the orthographic scheme of the interactive activation model does not represent diacritical letters. As a result, the application of this model to most Roman-based orthographies is a limitation in itself. To model the processing of diacritical letters, the orthographic scheme of the interactive activation model—assuming a more sophisticated letter feature level than with the Rumelhart and Siple (1974) uppercase font—can be extended along several lines.

In orthographies like German, where vowel diacritics mark distinct phonemes (<a> /a/ vs. <ä> /ɛ/), diacritical letters require distinct units at the letter level (Hutzler et al., 2004; Perea et al., 2021; Ziegler et al., 2000). In other words, each diacritical letter has its own node at the letter level, separate from its non-diacritical counterpart (see the left panel of Figure 2). This step necessitates modifications at the feature level

(i.e., the diacritical marks should be represented by additional features for diacritical letters), at the letter level (i.e., each diacritical letter functions as an independent letter node, relevant to both phonological mapping and orthographic distinction), and in the connection weights between levels, specifically between the features and letter nodes to indicate the distinctiveness of diacritical letters. This may include inhibitory connections between diacritical and non-diacritical letters, as they may compete during recognition, similar to what occurs with other visually similar letters (e.g., <i> and <j>, <n> and <m>).

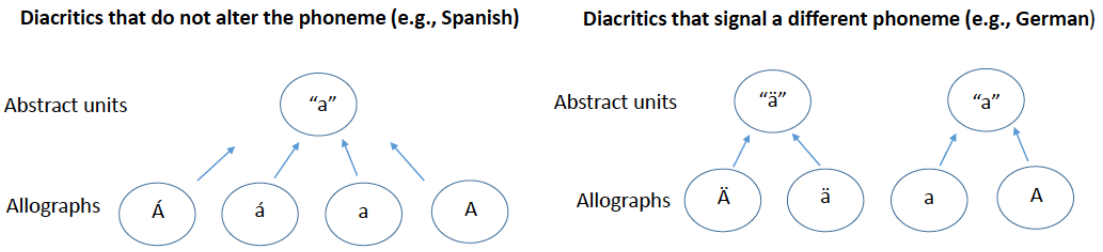

Figure 2. Illustration of how diacritics are modeled based on their function in the language.

In contrast, in those scenarios in which diacritical marks were purely suprasegmental (e.g., information about lexical stress) rather than indicating phoneme identity, as occurs with the accent marks in Spanish vowels, the abstract representations of diacritical letters could be shared by their non-diacritical partners (Perea et al., 2022). Under these circumstances at the feature level processing, the diacritical marks might be detected as noise at the feature level, without making a clear contribution to letter identification. In addition, at the letter level representation, diacritical and non-diacritical letters would share the same letter nodes (see right panel of Figure 2). This view is further supported by the fact that there is a minimal cost of processing when diacritics are removed or added in Spanish (Labusch et al.,

2022; Marcet & Perea, 2022; see also Duñabeitia et al., 2023, for evidence of a minimal impact of non-existing diacritics, as in <rëâdīńg>). Nonetheless, even in this scenario, we need to pay attention that diacritics are likely to be encoded and stored in the word identification system and they may play a role, especially for L1 or L2 learners of the script.

Notably, there are several potential challenges associated with extending computational models of visual word recognition to include diacritics. The first issue concerns the role of visual similarity: diacritical letters bear a strong visual resemblance to their non-diacritical counterparts, making them potentially confusable. Models need to be sensitive to visual similarity when discriminating between such letters—a similar case applies to letters like <i> and <j> or <C> and <G>. The second issue relates to perceptual asymmetries: there is a greater processing cost for adding a diacritic to a non-diacritical word than for omitting a diacritic in a diacritical word (Labusch et al., 2022, 2023). This asymmetry needs to be reflected in model mechanisms, either through differential weighting of features or activation thresholds. Third, language-specific phonological mappings should be considered: the function of diacritics varies by language. Therefore, models need to account for different types of grapheme-phoneme correspondences, which may sometimes require language-specific parameters or representational adjustments. Fourth, the frequency of diacritical letters in a given language can impact their processing. Diacritical letters are often less frequent than their non-diacritical counterparts (see Perea et al., 2020b, for discussion). High-frequency non-diacritical letters may form stronger representations in recognition systems, influencing activation levels.

Another important issue is how to incorporate diacritical letters in other scripts. As noted earlier, psycholinguistic research often reflects an Anglocentric bias (Share, 2008). In this chapter, we acknowledge a related limitation: a Latin-centric bias, as most findings and models are derived from languages that, while not English, also use the Roman script (e.g., German, French, Spanish). While insights from this Latin-centric research may not fully apply to non-Roman scripts, one could argue that the scenario in non-Roman scripts likely depends on the role of diacritics in each language.

To develop a comprehensive model of visual word recognition across most languages, it is crucial to extend current approaches to include letters with diacritics in their appearance to model reading processes more realistically. For the family of interactive activation models, this would require representing the letters with diacritical marks as distinct units, adjusting the activation dynamics, and language-specific parameters. Further research should refine these models through psycholinguistic experiments and by extending modeling to a variety of scripts and orthographies. Additionally, it is important to remember that this exploration can benefit from the “letter spirit' concepts proposed by Hofstadter (1985) (e.g., “What are the letters <a> and <á>?”; see Marcet et al., 2020).

4.4 Conclusions

Diacritics are not merely decorative marks but a basic building block of the orthography of most languages. Thus, their development illustrates how dynamic writing systems can be in adapting to the phonological system of their languages. Diacritical marks enhance written clarity, aid pronunciation, and help prevent confusion between homographs. Despite appearing minor to some linguists (Wells,

2001), their presence has important implications for language use and literacy. Understanding the processing of diacritics is important, as it provides insight into the structure of languages, the evolution of writing systems, and the strategies used for learning a new script.

References

AlJassmi, M. A. & Perea, M. (2024). Visual similarity effects in the identification of Arabic letters: Evidence with masked priming. *Language and Cognition*. https://doi.org/10.1017/langcog.2024.20

Adelman, J. S. (2011). Letters in time and retinotopic space. *Psychological Review, 118*(4), 570–582. https://doi.org/10. 1037/a0024811

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*, 439–461. https://doi. org/10.3758/bf03214334

Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multiple-trace memory model for polysyllabic word reading. *Psychological Review, 105*(4), 678–723. https://doi.org/10. 1037/0033-295x.105.4.678-723

Bae, S., Lee, C. H., & Pae, H. K. (2024). Visual letter similarity effects in Korean word recognition: The role of distinctive strokes. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218241278600

Balota, D., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (A travel update). In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd Edition) (pp. 285–375). Academic Press.

Benyhe, A., Labusch, M., & Perea, M. (2023). Just a mark: Diacritic function does not play a role in the early stages of visual word recognition. *Psychonomic Bulletin & Review*, *30*, 1530–1538. https://doi.org/10.3758/s13423-022-02244-4

Boudelaa, S., & Marslen-Wilson, W. D. (2010). ARALEX: A lexical database for Modern Standard Arabic. *Behavior Research Methods, 42*, 481-487. https://doi.org/10.3758/BRM.42.2.481

Chauncey, K., Holcomb, P. J., & Grainger, J. (2008). Effects of stimulus font and size on masked repetition priming: An event-related potentials (ERP) investigation. *Language and Cognitive Processes, 23,* 183-200. https://doi.org/10.1080/01690960701579839

Chetail, F., & Boursain, E. (2019). Shared or separated representations for letters with diacritics? *Psychonomic Bulletin & Review, 26*, 347–352. https://doi.org/10.3758/s13423-018-1503-0

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256. https://doi.org/10.1037/0033-295x.108.1.204

Comrie, B. (1996). Languages of Eastern and Southern Europe. In P. T. Daniels and W. Bright (Eds.), *The world's writing systems* (pp. 663 – 689). Oxford University Press.

Conrad, M., Tamm, S., Carreiras, M., & Jacobs, A. M. (2010). Simulating syllable frequency effects within an interactive activation framework. *European Journal of Cognitive Psychology, 22*, 861–893. https://doi.org/10.1080/09541440903356777

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review, 117*, 713–758. https://doi.org/10.1037/a0019 738

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences, 9*, 335–341. https://doi.org/10.1016/j.tics.2005.05.004

Duñabeitia, J. A., Perea. M., & Labusch, M. (2023). Rëâdīńg wõrdš wîth ōrńåmêńtš: is there a cost? *Frontiers in Psychology, 14*, 1168471. https://doi.org/10.3389/fpsyg.2023.1168471

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 680–698. https://doi.org/10.1037/0278-7393.10.4.680

Gómez, P., Marcet, A., & Perea, M. (2021). Are better young readers more likely to confuse their mother with their mohter? *Quarterly Journal of Experimental Psychology, 74*(9), 1542–1552. https://doi.org/10.1177/17470218211012960

Grainger, J. (2024). Letters, words, sentences, and reading. *Journal of Cognition, 7*(1), 66. https://doi.org/10.5334/joc.396.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*(3), 518–565. https://doi.org/10. 1037/0033-295x.103.3.518

Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: From pixels to pandemonium. *Trends in Cognitive Sciences, 12*, 381–387. https://doi.org/10.1016/j.tics.2008.06.006

Hofstadter, D. (1985). *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Basic Books.

Hutzler, F., Ziegler, J. C., Perry, C., Wimmer, H., & Zorzi, M. (2004). Do current connectionist learning models account for reading development in different

languages? *Cognition, 91*, 273–296. https://doi.org/10.1016/j.cognition.2003.09.006

Kinoshita, S., Yu, L., Verdonschot, R. G., & Norris, D. (2021). Letter identity and visual similarity in the processing of diacritic letters. *Memory & Cognition, 49,* 815–825. https://doi.org/10.3758/s13421-020-01125-2

Labusch, M., Gómez, P., & Perea, M. (2022). Does adding an accent mark hinder lexical access? Evidence from Spanish. *Journal of Cultural Cognitive Science, 6,* 219–228. https://doi.org/10.1007/s41809-022-00104-0

Labusch, M., Massol, S., Marcet, A., & Perea, M. (2023). Are goats chèvres, chévres, chēvres, and chevres? Unveiling the orthographic code of diacritical vowels. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49,* 301–319. https://doi.org/10.1037/xlm0001212

Marcet, A., & Perea, M. (2017). Is nevtral NEUTRAL? Visual similarity effects in the early phases of written-word recognition. *Psychonomic Bulletin and Review, 24*, 1180–1185. https://doi.org/10.3758/s13423-016-1180-9

Marcet, A., & Perea, M. (2018). Can I order a burger at rnacdonalds. com? Visual similarity effects of multi-letter combinations at the early stages of word recognition*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(5), 699. http://dx.doi.org/10.1037/xlm0000477

Marcet, A., & Perea, M. (2022). Does omitting the accent mark in a word affect sentence reading? Evidence from Spanish. *Quarterly Journal of Experimental Psychology*, *75*, 148–155. https://doi.org/10.1177/17470218211044694

Marcet, A., Fernández-López, M., Baciero, A., Sesé, A., & Perea, M. (2022). What are the letters e and é in a language with vowel reduction? The case of Catalan.

Applied Psycholinguistics, 43, 193–210. https://doi.org/10.1017/S0142716421000497

Marcet, A., Fernández-López, M., Labusch, M., & Perea, M. (2021). The omission of accent marks does not hinder word recognition: Evidence from Spanish. *Frontiers in Psychology*, *12*, 794923. https://doi.org/10.3389/fpsyg.2021.794923

Marcet, A., Ghukasyan, H., Fernández-López, M., & Perea, M. (2020). Jalapeno or Jalapeño: Do diacritics in consonant letters modulate visual similarity effects during word recognition? *Applied Psycholinguistics*, *41*, 579–593. https://doi.org/10.1017/S0142716420000090

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*, 375–407. https://doi.org/10.1037//0033-295x.88.5.375

Perea, M., Abu Mallouh, R., Mohammed, A., Khalifa, B., & Carreiras, M. (2016). Do diacritical marks play a role at the early stages of word recognition in Arabic? *Frontiers in Psychology, 7*, 1255. https://doi.org/10.3389/fpsyg.2016.01255

Perea, M., Abu Mallouh, R., Mohammed, A., Khalifa, B., & Carreiras, M. (2018). Does visual letter similarity modulate masked form priming in young readers of Arabic? *Journal of Experimental Child Psychology, 169*, 110–117. https://doi.org/10.1016/j.jecp.2017.12.004

Perea, M., Baciero, A., & Marcet, A. (2021). Does a mark make a difference? Visual similarity effects with accented vowels. *Psychological Research, 85*, 2279–2290. https://doi.org/10.1007/s00426-020-01405-1

Perea, M., Fernández-López, M., & Marcet, A. (2020a). Does CaSe-MiXinG disrupt the access to lexico-semantic information? *Psychological Research. 84*, 981–989. DOI: 10.1007/s00426-018-1111-7

Perea, M., Fernández-López, M., & Marcet, A. (2020b). What is the letter é? *Scientific Studies of Reading, 24*, 434–443. https://doi.org/10.1080/10888438.2019.1689570

Perea, M., Gomez, P., & Baciero, A. (2023). Do diacritics entail an early processing cost in the absence of abstract representations? Evidence from masked priming in English. *Language and Speech, 66,* 105–117. https://doi.org/10.1177/00238309221078321

Perea, M., Hyönä, J., & Marcet, A. (2022a). Does vowel harmony affect visual word recognition? Evidence from Finnish. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*, 2004–2014. https://doi.org/10.1037/xlm0000907

Perea, M., Labusch, M., & Marcet, A. (2022b). How are words with diacritical vowels represented in the mental lexicon? Evidence from Spanish and German. *Language, Cognition, and Neuroscience, 37*, 457–468. https://doi.org/10.1080/23273798.2021.1985536

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114*(2), 273–315. https://doi.org/10.1037/0033-295x.114.2. 273

Protopapas, A., & Gerakaki, S. (2009). Development of Processing Stress Diacritics in Reading Greek. *Scientific Studies of Reading, 13*(6), 453–483. https://doi.org/10.1080/10888430903034788

Reichle, E. D. (2020). *Computational models of reading: A handbook*. Oxford University Press.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89*(1), 60-94. https://doi.org/10.1037/0033-295x.88.5.375

Rumelhart, D. E., & Siple, P. (1974). Process of recognizing tachistoscopically presented words. *Psychological Review, 81*(2), 99-118. https://doi.org/10.1037/h0036117

Schwab, S. (2015). Accent mark and visual word recognition in Spanish. *Loquens, 2,* e018. https://doi.org/10.3989/loquens.2015.018

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin, 134*(4), 584–615. https://doi.org/10.1037/0033-2909.134.4.584

Snell, J. (2024). PONG: A computational model of visual word recognition through bi-hemispheric activation. *Psychological Review*. https://doi.org/10.1037/rev0000461

Wells, J. C. (2001). Orthographic diacritics and multilingual computing. *Language Problems and Language Planning, 24*, 249–272. https://doi.org/10.1075/lplp.24.3.04wel

Wiley, R. W., Wilson, C., & Rapp, B. (2016). The effects of alphabet and expertise on letter perception. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 1186–1203. https://doi.org/10.1037/xhp0000213

Winskel, H., & Perea, M. (2014). Does tonal information affect the early stages of visual-word processing in Thai? *Quarterly Journal of Experimental Psychology, 67*, 209–219. https://doi.org/10.1080/17470218.2013.813054

Ziegler, J. C., Perry, C., & Coltheart, M. (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology, 12,* 413–430. https://doi.org/10.1080/09541440050114570