What Is a Written Word?
And if So, How Many?

Martin Evertz-Rittich

Abstract. The linguistic unit *word* seems to be an intuitive notion for language users. However, linguists have failed so far to provide a uniform definition of that unit. Instead, there are definitions pertaining to different subsystems of language. In this paper, we will discuss how we can define the unit word in writing. We will start by examining definitions of the graphematic word in alphabetical writing systems such as German and English. We will then discuss how the written word relates to other suprasegmental units in writing systems, such as the syllable and the foot, and to which spoken unit or units a written word corresponds to. Finally, we will show that the discussed definitions of the graphematic word are not employable universally since in alphabetical writing systems definitions of the graphematic word pertain to interword spacing. By examining the Chinese and Japanese writing systems as examples, we will try to explain why these writing systems do not mark words by spaces and discuss whether there are graphematic words in these writing systems. Based on these considerations we will provide a tentative universal definition of graphematic words.

1. Introduction

Although the notion *word* seems to be an intuitive unit for language users—it might even be "the most basic of all linguistic units" (Taylor, 2015, p. 1)—it is a notoriously elusive concept in linguistics. This is due to the various criteria of wordhood in each linguistic subsystem, which often contradict each other. For instance, a phonological word, which (among other criteria) must exhibit exactly one primary stress, is not the same as a syntactic word, which (among other criteria) is moveable in a sentence (cf. *fish and chips* are three syntactical words but two phonological ones {fish and}{chips} with an unstressed *and*, cf. ibid., p. 7). Moreover, the criteria to identify a word in most subsystems of language are

Martin Evertz-Rittich 0000-0001-5434-6267
University of Cologne, Albertus Magnus Platz, 50923 Köln, Germany
E-mail: martin.evertz@uni-koeln.de

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2020. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 4.
Fluxus Editions, Brest, 2021, pp. 25–45. https://doi.org/10.36824/2020-graf-ever
ISBN: 978-2-9570549-6-1, e-ISBN: 978-2-9570549-8-5

often quite subtle and sometimes not even unambiguous (cf., e.g., the criteria for wordhood in semantics).

When it comes to the written word, however, things seem to be quite easy. Most often, the graphematic[1] word is defined as a string of letters bordered by spaces. And that seems to be the only noteworthy thing about that linguistic unit.

In this paper, I will show that there is more to the graphematic word. I will begin with the seemingly easy definition of the graphematic word and show that it is actually quite problematic. I will discuss the definition of the graphematic word in alphabetical writing systems, such as the writing systems of English and German, and show that the definitions found in the literature are insufficient. Based on typographic considerations, I present a promising alternative. In the next part, I will discuss the role of the graphematic word in the graphematic hierarchy and which properties can be derived from it. After that I will discuss the correspondence of the graphematic word to units in spoken language, such as the phonological and syntactical word (see above). Lastly, I will have a look at two writing systems that do not mark graphematic words by inter-word spacing: Japanese and Chinese. I will discuss why this is the case and whether there are graphematic words in these writing systems at all. In the conclusion, I will revisit the definition of the graphematic word presented in section 2 in light of the findings in section 5.

2. Definitions in Alphabetical Writing Systems

We will start our endeavor by examining definitions of graphematic words in alphabetical writing systems such as English or German. Probably the simplest definition is the one provided in (1).

(1) A graphematic word is a string of graphemes that is bordered by spaces and may not be interrupted by spaces.

This kind of definition is quite common in the literature (e.g., Coulmas, 1996, p. 550; Jacobs, 2005, p. 22; Fuhrhop, 2008, 193f). This definition seems to be intuitively correct and for most linguistic approaches—even grapho-linguistic ones—this definition suffices (cf., e.g., Evertz, 2018, p. 21). However, closer examination reveals that it is indeed problematic.

[1] In this paper, I will use the notion *graphematic* when conferring to a writing system. I refrain from using the term *orthographic* in this context since the orthography of a given writing system is the conventionalized spelling of that writing system and thus a subset.

But before we can begin discussing the definition, the terms within it must be clarified. In alphabetical writing systems, there are two traditions of defining the notion *grapheme*:

- A grapheme is a written unit that corresponds to exactly one phoneme (e.g., Wiese 2007).
- A grapheme is the smallest contrastive unit within a given writing system (e.g., Henderson, 1985, Kohrt, 1985, Eisenberg, 2006, Rogers, 2005).

While the first one defines the grapheme by its correspondence to phonological units, the second definition pertains to the distribution of the grapheme and thus is independent of phonology. The second definition closely corresponds to its counterpart in phonology, the definition of the *phoneme*. That entails that the grapheme, just like the phoneme, can be identified by minimal pair analyses.

The other term in the definition in (1) is the notion *space*. According to Bredel (2008, 31–32; 2011, 19–20) we can imagine the writing space as a threefold structure consisting of segmental slots, linear slots and two-dimensional slots, cf. Fig. 1.

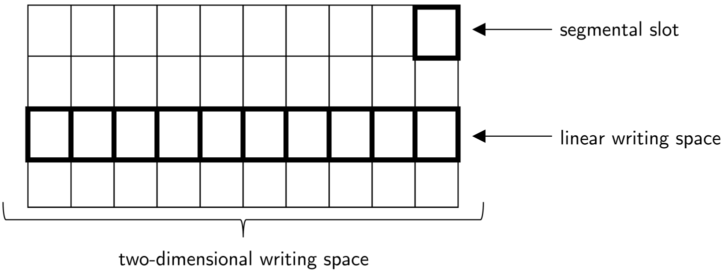

FIGURE 1. Writing Space (Bredel, 2011, p. 31; my translation)

Segmental slots are spaces that can be filled by certain graphic elements, e.g., letters. Linear writing spaces are horizontally oriented strings of slots. A two-dimensional writing space is a vertically oriented sequence of linear writing spaces (cf. Bredel, 2008, p. 19). A space according to (1) can be defined as an empty segmental slot.

Now that the terms in (1) are reasonably well clarified, we can have a closer look at this definition. Consider the examples in (2).

(2) <you.>, <you?>, <you!>
<Smiths'> (e.g., in *the Smiths' house*), <mother-in-law>

Let us start with the examples in (2a). According to the definition in (1), a word starts and ends with a grapheme. The examples in (2a), however, end in punctuation marks. These are not graphemes—regardless which definition of grapheme we employ: punctuation marks do not correspond to phonemes and they are not contrastive on the word level.

Thus, a word like <you> corresponds to the definition in (1), the examples in (2a), however, do not because they end in a punctuation mark.

If the definition in (1) is understood as being exhaustive (only those entities described in the definition qualify as graphematic words), the examples in (2a) are no graphematic words. But if they are not, what are they? If the definition in (1) is not exhaustive, it is not complete and additionally, the question arises, if the examples in (2a) are one or more words.

Similar problems arise with the examples in (2b). The word-status of <Smiths'> is unclear as is the question whether a hyphenated word (?) like <mother-in-law> constitutes one or more graphematic words.

One alternative to the definition in (1) is proposed by Zifonun, Hoffmann, and Strecker (1997, p. 259), my translation:

(3) A graphematic word is a string of graphemes that is preceded by a space and may not be interrupted by spaces.

This definition only seemingly solves the problems we have encountered so far. The examples in (2a) constitute according to this definition exactly one graphematic word, <you>, because the "string of graphemes" is interrupted by a punctuation mark in each case. The same is true for the first example in (2b). This string of graphemes is interrupted by the apostrophe. The case of the second example in (2b) is more complicated, however. According to the definition in (3), <mother-in-law> constitutes exactly one graphematic word: <mother>. The status of <-in-law> is unclear.

Moreover, there are examples like in (4) that do not only end but also begin with punctuation marks.

(4) <"you">, <(you)>, <¿tú?> (Span.)

Thus, the definition in (3) is also problematic.

The solution I propose is based on typographic considerations by Bredel (2008; 2011). Based on the model of writing space (cf. Fig. 1) we can distinguish between two classes of punctuation marks and graphemes: *fillers* and *clitics*. Fillers can independently fill a segmental slot whereas clitics need the support of a filler.

Bredel (2008) proposes two criteria by which fillers and clitics can be distinguished. The first one is symmetry. One element is called symmetric, if elements of the same class can stand adjacent to the left and right side of that element. Fillers are symmetric, clitics are not. The second criterium is the ability of an element to appear at the beginning and the end of a line. Fillers can appear at the end and the beginning of a line, clitics cannot.

According to Bredel (2011, pp. 20–23) letters, numbers, apostrophes and hyphens are fillers; periods, colons, semi-colons, commas, brackets, question marks, quotation marks and exclamation marks are clitics.

Based on this distinction, we propose the following definition of graphematic words in alphabetical writing systems such as German and English (Evertz, 2016a, pp. 391–392); based on works of Bredel; my translation):

(5) A graphematic word is a sequence of slot-filler-pairs surrounded by empty slots in which at least one filler must be a letter.

The supplement to the definition (at least one filler being a letter) was added to exclude numbers from the scope of the definition. Let us examine one of our examples in light of this definition, cf. Fig. 2.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	m	o	t	h	e	r	-	i	n	-	l	a	w!	

FIGURE 2. Slot-filler-pairs of <mother-in-law!>

In Fig. 2 there are 15 segmentals slots. Slots 2 to 14 are occupied, slots 1 and 15 are empty. Slots 2 to 7, 9 to 10 and 12 to 13 are occupied by one letter each, slot 14 is occupied by a letter and a punctuation mark. Slots 8 and 11 are each occupied by one non-letter filler. Thus, <mother-in-law> meets all requirements for a graphematic word according to the definition in (5).

The consequence of the definition in in (5) is that we can distinguish between the graphematic word proper and its surface form. Clitics are only part of the graphematic surface whereas fillers are part of the graphematic surface *and* of the graphematic word proper. This is true for all fillers: letters and non-letters (cf. ibid., pp. 391–392).

In the case of the examples in (2a) and (4), the graphematic word proper consists of the fillers: <you>. The clitics (in these cases the punction marks) are part of the graphematic surface. Thus, the examples in (2a) and (4) are graphematic surface forms of exactly one graphematic word (cf. ibid., pp. 391–392).

The examples in (2b) consist exclusively of fillers. This means that all characters (letters and non-letters alike) make up the graphematic word proper. The non-letter fillers are part of the graphematic word since they have important roles within it. In the case of <mother-in-law> (cf. Fig. 2), the non-letter fillers indicate that the morphological processing is not completed after <mother> and <in> but that everything between the empty slots must be processed as whole (Evertz 216a, 391). In the case of <Smiths'> as in *the Smiths' house*, the apostrophe indicates a zero morpheme (Bunčić, 2004, p. 190). A consequence is, however, that <Smiths'> and <Smith> are two different graphematic words.

This definition is very promising for writing systems such as English and German. We will see however, that it is a poor candidate for a universal definition, cf. section 5.

3. Properties of Graphematic Words

The graphematic word is a unit in writing systems that issuprasegmental, i.e., it is larger than a single segment. It is not the only suprasegmental unit in alphabetical writing systems. The graphematic syllable is well-established in psycho- and grapholinguistic literature (e.g., Butt and Eisenberg, 1990; Domahs, Bleser, and Eisenberg, 2001; Eisenberg, 2006; Primus, 2003; Rollings, 2004; Roubah and Taft, 2001; Weingarten, 2004) and more recently, the graphematic foot gained attention (Evertz, 2016a,b; 2018; 2019; Evertz and Primus, 2013; Fuhrhop and Peters, 2013; Primus, 2010; Ryan, 2018). With these units it is possible to constitute a graphematic counterpart of the phonological hierarchy, cf. Fig. 3.

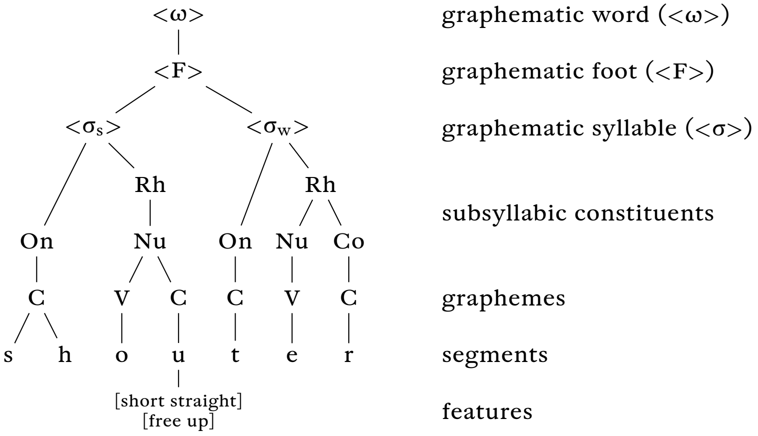

FIGURE 3. The graphematic hierarchy (Evertz, 2018; Evertz and Primus, 2013

This hierarchy is—just as its phonological counterpart—accompanied by the *Strict Layer Hypothesis* (Nespor and Vogel, 1986). This hypothesis states in its strong form that each unit of a non-terminal category is composed of one or more units of the immediately lower category. The second part of the Strict Layer Hypothesis states that a unit of a given level of the hierarchy is exhaustively contained in the superordinate unit of which it is part (ibid., p. 7). Previous work showed that this hypothesis also holds in graphematics, although it seems that the first principle is violable in case of so called extrametrical syllables (cf. Evertz, 2018).

A consequence of these considerations are that:

- a graphematic word consists of at least one graphematic foot and
- a graphematic foot consists of at least one graphematic syllable.

Since larger units in a hierarchy are made up of the immediately smaller units, the larger units inherit traits of the smaller units. For instance, if a syllable must adhere to certain well-formedness requirements and if a foot is constituted by syllables, the syllables of the foot must adhere to the very the same requirements. The same is true on every level of the hierarchy. This means that a graphematic word must adhere to well-formedness requirements of graphematic feet and graphematic syllables.

This relationship can be exemplified by so called *minimal words* (Evertz, 2016b). Consider following examples:

(6) in/inn, oh/owe, no/know, by/bye/buy, so/sew, to/two, we/wee, or/ore/oar, be/bee, I/aye/eye

The pairs or triplets in (6) are homophones. Interesting is that function words can obviously be shorter than content words. This can be described by the so called *three-letter-rule* (e.g., Cook, 2004, p. 57):

(7) Content words must have more than two letters.

The existence of a minimality restriction like the three-letter-rule can be explained with the help of the graphematic hierarchy.

Just like in phonology, we can expect that function words behave differently than content words. For instance, while content words always constitute phonological words, which exhibit exactly one prime stress, function words can be unstressed. In phonology, this can be described by following constraint:

(8) LEXWD = PRDWD: Every lexical word corresponds to a prosodic word (ibid., p. 101).

Let us assume that this constraint also holds for writing systems:

(9) Every lexical word corresponds to a well-formed graphematic word.

The difference in the pair and triples in (6) can now be explained by the well-formedness constraints the graphematic word inherits from the smaller units in the hierarchy.

In phonology, feet must conform to a certain well-formedness constraint, called *foot-binarity* (McCarthy and Prince, 1995, pp. 320–324):

(10) Foot-Binarity: Feet are binary at a syllabic or moraic level of analysis.

This means that a well-formed phonological foot must consist of two syllables or one heavy syllable. Evertz (2016b; 2018) shows that a similar constraint holds for graphematics. A graphematic foot must consist of either one heavy graphematic syllable or two graphematic syllables (of any weight). Whether a graphematic syllable is heavy or light depends on its syllabic structure. In order to be heavy, a graphematic syllable must have a rhyme that dominates at least two segments and in total the syllable must consist of at least three segments (Evertz, 2016b, p. 208; see the fist syllable in Fig. 3 as an example of a heavy graphematic syllable).

Although the three-letter-rule is not wrong, the explanation provided here is superior in explanatory strength. Moreover it is empirically superior. If having three letters was the only restriction for graphematic words, there should be more words like <gnu>, which end in a single vowel letter but still consist of three letters. Words of this type, however, are quite rare (cf. ibid.).[2]

Even the fact that content words have at least one vowel letter can be derived from the graphematic hierarchy: A graphematic word consists of at least one foot. A graphematic foot consists of at least one graphematic syllable. And a graphematic syllable must have a core dominating at least one vowel letter (e.g., Evertz, 2018; Fuhrhop and Peters, 2013; Primus, 2003).

There are, however, exceptions to the well-formedness constraints described here. Graphematic words that systematically violate these constraints are abbreviations:

[2] Evertz (2016b, p. 193) reports that only 20.4% of monosyllabic phonological words ending in a vowel are written as a monosyllabic graphematic word ending in a single vowel letter. Of these 20.4%, 4.6% are function words, 9.2% are loanwords, interjections or abbreviations, leaving a rest of 6.7%. Monosyllabic phonological words ending in a vowel are rather coded with the help of so called mute letters like in *blow, bee, high*. Evertz (ibid., pp. 207–208) argues that these mute letters add graphematic weight in order to meet the weight restriction for graphematic feet.

(11) Examples for graphematic words violating well-formedness constraints:
– ill-formed graphematic syllables: Mr., Mrs., vs., Dr.
– ill-formed graphematic feet: BA, MA, no.

The examples in (11a) violate the constraint that the cores of graphematic syllables dominate a vowel letter. The examples in (11b) violate the constraint that graphematic feet need to have a minimal weight.

Those words that violate well-formedness constraint are marked by special orthographic devices like dots or all-caps. We may thus describe such abbreviations as untypical and marked graphematic words (Evertz, 2016a, p. 393).

4. Relations to Phonological Units

After having discussed the definition of the graphematic word and some of its properties, let us now try to discuss the relationship of the graphematic word to other word-like units.

Let us begin with the *phonological word*. It is quite obvious that the phonological word and the graphematic word are not congruent. A phonological word is a linguistic unit that consists of at least one phonological foot and exhibits exactly one primary stress. Within phonological words, syllable boundaries are drawn according to *onset maximization* (assign as many intervocalic consonants to the onset as possible (in accordance with the phonotactical constraints of a language); e.g., Giegerich, 1992, p. 170). For instance, *tomato* constitutes exactly one phonological word. There are several potential ways to divide the word into syllables, e.g., **tom.at.o* vs. *to.ma.to*. Only the second way conforms to the onset maximization principle. However, onset maximization does not incur, if a border of a phonological word is interfering.

One example for that is the German compound *Tierart* 'animal species'. According to onset maximization, the intervocalic consonant /r/ should be the onset of the second syllable. However, this syllabification is ungrammatical: *[tiː.ʁaːɐ̯t]. Instead, the word is syllabified like this: [tiːɐ̯.ʔaːɐ̯t]. Thus, we can conclude that a phonological word border is interfering with onset maximization. In other words, *Tierart* consists of two phonological words: {*Tier*}{*art*}. However, it is realized graphically as one graphematic word <Tierart>. Therefore it seems that in German, the phonological word and the graphematic word are incongruent.

An example of the incongruity of phonological and graphematic words in English was mentioned in the first section: *fish and chips*. While this phrase consists of three graphematic words, it consists of only two phonological words: {fɪʃn̩}{ʧɪps} (Taylor, 2015, p. 7).

A *morphological word* can be described a an entity that inflects uniformly (Wurzel, 2000, p. 36) and is constituted by word building rules (Jacobs, 2005). Thus, our example *Tierart* is a morphological word since it is constituted according to the composition rules of German and is inflecting uniformly: *Tierarten* (Pl.) vs. **Tierearten*. Fuhrhop (2008, p. 224) comes to the conclusion that the morphological word is congruent with the graphematic word in German.[3]

A *syntactic word* can be defined as a syntactically free form that is commonly designated X^0 in *generative grammar* (cf. Gallmann, 1999). This entails that a syntactical word is permutable in a sentence and may not be interrupted by linguistic material. Gallmann (ibid.) and Fuhrhop (2008) come to the conclusion that the syntactic word and the graphematic word are almost congruent[4] in German.

From a writer's perspective the congruity of graphematic words with syntactical and morphological words means that phrases must be realized as single graphematic words with empty slots in between. Complex morphological words, however, must be realized as one graphematic word without empty slots in between. Conversely, from a reader's perspective this means that a slot-filler-sequence without spaces must be interpreted morphologically and slot-filler-sequences with spaces must be interpreted syntactically. This can be exemplified by *wohlgeraten* 'great, outstanding' vs. *wohl geraten* 'probably guessed'. Because there are no empty slots in *wohlgeraten*, it must be interpreted as one graphematic word and therefore as one morphological word. And because there is an empty slot in *wohl geraten*, this expression must be interpreted as two graphematic words and therefore two syntactic words, a phrase in this case.

The case for English is not as straightforward as in German. This is due to the fact that there is a considerable stylistic freedom in the spelling of compound words. For instance, the website *Wiktionary* lists three spellings of *secondhand*: <secondhand>, <second-hand> and <second hand>. However, as the same website points out, <secondhand> and <second-hand> "may be preferred spellings for the adjective meaning 'not new', to avoid confusion with the noun 'second hand' referring to the hand of a clock or watch."[5]. This means that

[3] Whether there are exceptions to the congruity of morphological and graphematic words is debatable. Wurzel (2000, p. 37) points to the case of *(mit seiner) Langenweile* '(with his) boredom (Dative)' a variant of *Langeweile*. This (not too common) variant may suggest that *Langeweile* is actually consisting of two morphological words but one graphematic word.

[4] Examples include particle verbs like *anfangen* 'to begin' in sentences like *er fängt an zu schreiben* 'he starts writing'.

[5] `https://en.wiktionary.org/wiki/second_hand#English`, retrieved August 21st, 2020.

spellings without empty slots are quite clearly interpreted as one morphological word while spellings with empty slots can have ambiguous readings. Evertz (2016a, p. 394) points to the example *old furniture dealer*: an <old-furniture dealer> is a dealer of old furniture, an <old furniture-dealer> is a furniture dealer who is elderly.

Thus it seems that just like in the German writing system, the graphematic word is congruent with the morphological and syntactical word in English although the English writing system allows more variation in writing compound words.

5. Graphematic Words Without Spaces?

So far we examined the graphematic word in English and German as examples of alphabetical writing systems that use empty slots to mark the beginning and end of graphematic words. However, there are writing systems, alphabetical and non-alphabetical, that do not use empty slots in that way. In this section, we will have a look at two examples and discuss why in these cases there are no empty slots and whether we still can find reasons to assume that the graphematic word is a relevant unit in these writing system.

5.1. The Case of Japanese

The Japanese writing system (JWS) is regarded as one the most complex writing systems in the world (e.g., Joyce, 2011). Sproat (2010, p. 47) for instance writes that "Japanese is a complex system, certainly the most complex writing system in use today and a contender for the title of the most complex system ever." The reason for this consensus regarding its complexity is the multitude of scripts employed in the Japanese writing system. In the contemporary JWS there are five separate scripts: morphographic *kanji*, the mora-based (Ratcliffe, 2001) scripts *hiragana* and *katakana*, the phonemic Roman alphabet *rōmaji* and Arabic numerals (e.g., Joyce and Masuda, 2018, p. 182).

The different scripts are used for different purposes. Kanji are generally used to represent native and Sino-Japanese content words like nouns, the stem of verbs etc. (ibid., p. 184). For instance, the compound 日本語 *nihongo* 'Japanese' consists of three kanji 日本 'Japan' and 語 'language'.Hiragana, on the other hand, generally represent function words such as auxiliaries, and inflectional endings (ibid., p. 184). In this use they are referred to as 送り仮名 *okurigana* 'accompanying letters'. An example for okurigana are the hiragana following the kanji in 見る *miru* '(to) see' vs. 見た *mita* 'saw'. Katakana are usually used to write non-Chinese loanwords, foreign names, animal and plant species names, onomatopoeic expressions, and for emphasis and as glosses. Rōmaji are similarly used to represent non-Japanese words and names, especially within advertising and mass media. And finally, Arabic numerals are used to represent numbers, particular in financial and scientific contexts (Joyce and Masuda, 2018, p. 184).

While on first sight this multitude of different scripts might seem confusing, it can actually be beneficial for readers as they enable them to distinguish lexical content from grammatical elements (Joyce and Masuda, 2016). This is because of the visual distinctiveness of the three scripts the JWS mainly uses. First, kanji are visually salient because of their complexity. In contrast to hiragana and katakana, which are usually written with no more than six strokes (Kajii, Nazir, and Osaka, 2001, p. 2504), kanji can consist of up to 29 strokes with an average of 10.47 strokes (Joyce, Hodošček, and Nishina, 2012, p. 256; Joyce and Masuda, 2018, p. 186). Since these salient units usually represent lexical content, it can be identified at first glance. Second, hiragana are also easily identifiable: they consist of relatively few strokes, which tend to be curved, in contrast to katakana, which consist of more or less the same amount of strokes, which, however, tend to be straight. Thus, grammatical elements, which are usually represented by hiragana, are also quite easily identifiable. Reading experiments confirmed that readers can distinguish the three types of characters effortless, even in peripheral vision (Osaka, 1989; 1992). Given the foreignness in appearance of rōmaji and Arabic numerals, it is quite reasonable to assume that they too can be distinguished easily by readers of the JWS.

Let us demonstrate the interplay of the different scripts within the JWS, cf. the example in (12).

(12) Example for the interplay of different scripts in the JWS (Shibatani, 1990, p. 129)

花子	は	あの	ビル	で	働	いている	ＯＬ	です。
Hanako	wa	ano	biru	de	hatari-	i-te-i-ru	ooreu	desu
Hanako	Topic	that	building	at	work-	ing	OL	is

‘Hanako is an OL (office lady) working in that building’

Content words (in one case a verb stem) are represented by kanji (花子, 働), by katakana (ビル) or rōmaji (*OL*). Since in Japanese inflectional endings are following the stem, word beginnings coincide with characters that usually represent lexical content, especially kanji (cf. Rogers, 2005, p. 66). Thus, characters frequently appearing in the word beginning may serve as effective segmentation cues to signal word boundaries.

This points to the conclusion that graphematic words do not need to be explicitly marked by empty slots in Japanese, since the words are already marked *graphotactically*. This conclusion is supported by psycholinguistic findings. Sainio, Hyönä, Bingushi, and Bertram (2007) found that interword spacing facilitated Japanese readers—but only when they read a text composed of hiragana only. In normal Japanese texts, which mainly consist of kanji and hiragana, interword spacing did not facilitate reading.

5.2. The Case of Chinese

Like the Japanese writing system, the Chinese writing system (CWS) does not display empty slots between individual characters, which represent most likely a morpheme or a syllable, cf. (13).

(13) Example of a Chinese sentence without interword spacing
中国这几年的变化的确很大。

The sentence neither displays spacing between words or phrases nor does it display graphotactical cues to word boundaries like in the JWS. Yet there are linguistic units greater than single syllables, morphemes or characters. (14) provides a translation of the sentence in (13), in which syntactic words are separated.

(14) Translation of the sentence in (13) (Coulmas, 2003, p. 59)

中国	这	几	年	的	变化	的确	很	大。
Zhōngguó	zhè	jǐ	nián	de	biànhuà	díquè	hěn	dà
China	these	several	years	Gen	change	really	very	big

‘China underwent big changes during the past several years’

In the CWS, syntactic words can be written with one or more characters, as seen in (14). A word comprising two characters is not necessarily a compound word. For instance, in 蚯蚓 *qiūyǐn* ‘earthworm’ neither character represents a morpheme but both characters combined do (Chen, 1996, p. 46). An example for the difference between a phrase and a syntactic word written with two characters is the contrast between 红鸟 *hóngniǎo* ‘red bird’ and 红花 hónghuā ‘safflower’ (examples from Zhang, 1985, p. 64 as cited in Packard, 2000, p. 15). Notice that in both cases the first character is 红, which in isolation denotes ‘red’. In 红鸟, there are two syntactic words because both components can be substituted by nearly any adjective and any noun while it still retains its compositional meaning. In 红花, on the other hand, the idiomatic meaning gets lost by substituting one component (ibid., p. 15).

The Common Words in Contemporary Chinese Research Team (2008) analyzed a corpus consisting of 56,008 words and found that 6% of Chinese words are written with a single-character, 72% are 2-character words, 12% are written with 3 characters, and 10% are 4-character words; fewer than 0.3% of Chinese words are written with more than 4 characters. Analyzing the token frequencies, 70.1% of words are written with a single character, 27.1% are 2-character words, 1.9% are 3-character words, 0.8% are 4-character words, and 0.1% are words longer than 4 characters.

This means that 94% of words (types) are longer than one character and even by taking tokens into account, nearly 30% of words are still larger than a single character. This leads to the question why the CWS does not display empty slots between words and whether there is a graphematic counterpart to the syntactic word in Chinese.

One reason for the lack of interword spacing might lie in the development of the CWS. Classical Chinese was mostly monosyllabic and monomorphematic, thus words and characters were almost congruent (Hoosain, 1992, p. 119; Li, Zang, Liversedge, and Pollatsek, 2015, p. 232). Therefore, the writing system of Classical Chinese had simply no need for interword separation.

Packard (1998; 2000) mentions the fact that there was no term for the syntactic word in the Chinese language until the concept was imported from the West at the beginning of the twentieth century. This new term is called 词 cí 'syntactic word'. It describes a concept that is quite different from the older word that is still used in non-linguistic contexts when talking about word-like entities in Chinese, 字 zì, which can be translated as 'morpheme-syllable' or 'character' (Hoosain, 1992, p. 112).

A reason why interword spacing did not develop over time in the Chinese writing system (CWS) might be due to the linguistic features of contemporary Chinese. It is noteworthy that modern Chinese almost completely lacks inflection. Thus, unlike in the JWS, there is no need for a non-morphemic script for grammatical information in the CWS. Moreover, Hoosain (ibid., pp. 118–120) reports that morphemes in Chinese can be free or bound. However, there are degrees of freedom as the free-bound status of a morpheme can vary by context, register and dialect. Lastly, bound morphemes can appear before or after a free morpheme, unlike in many other languages which do only allow bound morphemes to either appear before or after a free morpheme (Chen, 1996, p. 46). According to Hoosain (1992, p. 120), these factors contribute to a "fluidity of word boundaries" in the mind of Chinese speakers. Thus, a distinction between morphemes and words in the CWS would not be appropriate. Packard (2000, pp. 17–18), however, disputes this argument. He argues that Chinese speakers might only be uncertain in their metalinguistic judgment but will have no problems in actual language usage.

As an interesting side note, Meng et al. (2019) compared the efficiency of deep learning-based Chinese natural language processing algorithms. They benchmarked neural word-based models which rely on word segmentation against neural character-based models which do not involve word segmentation in four tasks (language modeling, machine translation, sentence matching/paraphrase and text classification). They found that character-based models consistently outperformed word-based models.

While the linguistic argument of Hoosain (1992) is under dispute, there is however consensus about the average word length in Chinese. As reported further above, ca. 78% of word types and ca. 97% of word tokens are one or two characters in length. This leads Li, Zang, Liversedge, and Pollatsek (2015) to another interesting explanation why there is no interword spacing in the CWS: the variance in word length in Chinese is reduced relative to the word length variability in alphabetic languages. The number of potential sites within a character string at which word segmentation might occur is therefore significantly reduced in Chinese. Consequently, decisions about word boundaries might be less of a challenge in Chinese than in English (given English had no empty slots). Thus, word spacing may have been less of a necessity for efficient reading in Chinese (ibid., pp. 232–233).

These considerations are supported by psycholinguistic findings. The interspersing of spaces (or other highlighting) between syntactic words does not facilitate reading Chinese, but did not interfere with reading in adult readers as well (Bai et al., 2008; Inhoff, Liu, Wang, and Fu, 1997). Inserting a space after a word facilitates its processing but inserting a space before a word did not facilitate processing and in fact may even interfere with its integration into sentential meaning as indicated by total reading times (Li and Shen, 2013; Liu and Li, 2014).

To sum these considerations up: In classical Chinese, there was no need to introduce a delimiter of words since words and characters were almost congruent. In contemporary Chinese this is not the case. There is a considerable amount of syntactic words that are written with more than a single character. But because of linguistic features of the Chinese language which allow morphemes to occur relatively freely in different syntactical contexts and because of the relatively reduced word length variability in Chinese, it seems that the character is the central unit for reading Chinese.

Thus it seems that the graphematic word is simply not a relevant—or existing—unit in the CWS. This is an important insight for suprasegmental graphematics pertaining to the role of the graphematic hierarchy across languages. While the phonological counterpart of the graphematic hierarchy, the prosodic hierarchy, is assumed to be universal[6], the writing system of Chinese demonstrates that at least the graphe-

[6] But see, e.g., Schiering, Bickel, and Hildebrandt (2010), who question the universality of the phonological word and find evidence that there are more units within the prosodic hierarchy than assumed.

matic word is not a universal category of the graphematic hierarchy. This opens the debate whether all units within the graphematic hierarchy are universal and whether the graphematic hierarchy as a whole is universal across writing systems at all.

6. Conclusion

The definition that the graphematic word is a string of graphemes bordered by spaces, which is well-accepted in the literature, turns out to be problematic because it does not take the role of punctuation marks into account. A promising alternative to this definition is typography-based. In this definition a graphematic word is defined as a sequence of slot-filler pairs, in which at least one filler is a grapheme, bordered by empty slots. This definition has the benefit that it allows to distinguish between the graphematic surface and the graphematic word proper. Clitics belong to the graphematic surface of a word only.

The graphematic word is part of the graphematic hierarchy, the graphematic counterpart to the phonological hierarchy. Taking the strict layer hypothesis into account, it is possible to explain certain features of the graphematic word. Since graphematic words consist of graphematic feet, which in turn consist of graphematic syllables, the graphematic word inherits traits of the foot and the syllable. One example for such a trait is the fact that graphematic words must have at least one vowel letter: because graphematic syllables need to have a vowel letter in their core, a graphematic word needs to have at least one vowel letter as well. Another example provided in this paper is the minimal weight restriction for graphematic words. The existence of this restriction can be explained by a well-formedness constraint of graphematic feet stating that a foot must be binary in syllabic or moraic terms.

Examining the German and English writing systems, it seems that the graphematic word mainly corresponds to the morphological and syntactical word in spoken language. A graphematic word written with no empty slots in between is interpreted as one morphological unit in both writing systems. Empty slots on the other hand indicate distinct syntactical units in the German writing system. In the English writing system, there is a greater variety in writing compound words. The use of a hyphen (a filler according to the typographic considerations in section 5) or the avoidance of empty slots may however disambiguate unclear cases.

In some writing systems there are no empty slots between characters. However, it can be argued that there are graphematic words in the Japanese writing system, which are not marked by empty slots but by graphotactical means. In the Japanese writing system, hiragana are used to represent function words and inflectional endings while other scripts (especially kanji) are used to represent lexical information. Because lexical words usually start with a kanji character (or katakana or rōmaji), the beginning of a graphematic word can easily be spotted.

If we accept that graphematic words do exist in Japanese, which is a writing system without empty slots between words, the definition of graphematic words in (5) is not universal. A universal definition of graphematic words has to include that in some writing systems, graphotactical means are used to mark the borders of graphematic words.[7] This universal definition must therefore be quite broad and unspecific. Subdefinitions pertaining to certain writing systems or families of writing systems are needed to supplement this broad universal definition. The definition in (15) is a first tentative proposal.

(15) A graphematic word is a sequence of slot-filler pairs, in which at least one filler must be a basic unit of the given writing system.

1. This sequence is bordered by empty slots or
2. the beginning of that sequence is indicated by other graphotactical means (e.g., the change of scripts).

The term basic unit is a deliberately broad term to accommodate different types of writing systems. However, it might not be quite clear what the basic unit of a given writing system is. In case of the JWS, it is fair to say that the characters of kanji, hiragana and katakana are basic units of the writing system. But it is unclear whether the characters of rōmaji are belonging to this class. Furthermore, while the notion of empty slots is quite clear, the term "graphotactical means" is quite fuzzy as well. In both cases, writing system specific sub-definitions must be supplemented.

Another insight we gained from examining writing systems without empty spaces pertains to the graphematic hierarchy. In the Chinese writing system, words are neither marked by empty slots nor by other graphotactical means. Thus it seems that the graphematic word is not a relevant unit in the Chinese writing system. This is an interesting finding for suprasegmental graphematics. In suprasegmental phonology, it is claimed that all the units of the prosodic hierarchy are universal. In graphematics, however, it seems that this is not the case—at least for the graphematic word. Further typological investigations are needed to explore the role of the graphematic hierarchy in non-alphabetical writing systems.

[7] The Thai writing system may also be a candidate for a system marking its graphematic words by graphotactical means.

References

Bai, Xuejun et al. (2008). “Reading spaced and unspaced Chinese text: Evidence from eye movements.” In: *Journal of Experimental Psychology: Human Perception and Performance* 34, pp. 1277–1287.

Bredel, Ursula (2008). *Die Interpunktion des Deutschen. Ein kompositionelles System zur Online-Steuerung des Lesens*. Tübingen: Max Niemeyer.

——— (2011). *Interpunktion*. Heidelberg: Winter.

Bunčić, Daniel (2004). “The apostrophe: A neglected and misunderstood reading aid.” In: *Written Language and Literacy* 7.2, pp. 185–204.

Butt, Matthias and Peter Eisenberg (1990). “Schreibsilbe und Sprechsilbe.” In: *Zu einer Theorie der Orthographie*. Ed. by Christian Stetter. Tübingen: Niemeyer, pp. 33–64.

Chen, May Jane (1996). “An overview of the characteristics of the Chinese writing system.” In: *Asia Pacific Journal of Speech, Language and Hearing* 1.1, pp. 43–54.

Cook, Vivian (2004). *The English Writing System*. London: Routledge.

Coulmas, Florian (1996). *The Blackwell Encyclopedia of Writing Systems*. Oxford: Blackwell.

Domahs, Frank, Ria de Bleser, and Peter Eisenberg (2001). “Silbische Aspekte segmentalen Schreibens – neurolinguistische Evidenz.” In: *Linguistische Berichte* 185, pp. 13–30.

Eisenberg, Peter (2006). *Grundriß der deutschen Grammatik. Bd. 1: Das Wort*. 3rd ed. Stuttgart: J.B. Metzler.

Evertz, Martin (2016a). “Graphematischer Fuß und graphematisches Wort.” In: *Laut – Gebärde – Buchstabe*. Ed. by Beatrice Primus and Ulrike Domahs. Berlin/New York: De Gruyter, pp. 377–397.

——— (2016b). “Minimal graphematic words in English and German. Lexical evidence for a theory of graphematic feet.” In: *Written Language & Literacy* 19.2, pp. 189–211.

——— (2018). *Visual Prosody—The graphematic foot in English and German*. Berlin, New York: De Gruyter.

——— (2019). “The History of the Graphematic Foot in English and German.” In: *Graphemics in the 21st Century*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 27–40.

Evertz, Martin and Beatrice Primus (2013). “The graphematic foot in English and German.” In: *Writing Systems Research* 5.1, pp. 1–23.

Fuhrhop, Nanna (2008). “Das graphematische Wort (im Deutschen): Eine erste Annäherung.” In: *Zeitschrift für Sprachwissenschaft* 27, pp. 189–228.

Fuhrhop, Nanna and Joerg Peters (2013). *Einführung in die Phonologie und Graphematik*. Stuttgart: J. B. Metzler.

Gallmann, Peter (1999). “Wortbegriff und Nomen-Verb-Verbindungen.” In: *Zeitschrift für Sprachwissenschaft* 18.2, pp. 269–304.

Giegerich, Heinz J. (1992). *English phonology: An introduction*. Cambridge: Cambridge University Press.

Henderson, Leslie (1985). “On the use of the term ‘grapheme’.” In: *Language and Cognitive Processes* 1.2, pp. 135–148.

Hoosain, Rumjahn (1992). “Psychological reality of the word in Chinese.” In: *Advances in psychology 90: Language processing in Chinese*. Ed. by Hsuan-Chih Chen and Ovid J.L. Tzeng. Amsterdam: North-Holland, pp. 111–130.

Inhoff, Albrecht W. et al. (1997). “Use of spatial information during the reading of Chinese text.” In: *Cognitive research on Chinese language*. Jinan: Shan Dong Educational Publishing, pp. 296–329.

Jacobs, Joachim (2005). *Spatien. Zum System der Getrennt- und Zusammenschreibung im heutigen Deutsch*. Berlin: De Gruyter.

Joyce, Terry (2011). “The significance of the morphographic principle for the classification of writing systems.” In: *Written Language & Literacy* 14.1, pp. 58–81.

Joyce, Terry, Bor Hodošček, and Kikuko Nishina (2012). “Orthographic representation and variation within the Japanese writing system: Some corpus-based observations.” In: *Written Language & Literacy* 15.2, pp. 254–278.

Joyce, Terry and Hisashi Masuda (2016). “Just mixed up or a pretty neat idea? Some reflections on the multi-script nature of the Japanese writing system.” In: *Understanding writing systems: From core issues to implications for written language acquisition’—10th International Workshop on Written Language and Literacy*. Radboud University, Nijmegen.

——— (2018). “Introduction to the multi-script Japanese writing system and word processing.” In: *Writing Systems, Reading Processes, and Cross-Linguistic Influences: Reflections from the Chinese, Japanese and Korean Languages*. John Benjamins, pp. 179–200.

Kajii, Natsumi, Tatjana A. Nazir, and Naoyuki Osaka (2001). “Eye movement control in reading unspaced text: the case of theJapanese script.” In: *Vision Research* 41, pp. 2503–2510.

Kohrt, Manfred (1985). *Problemgeschichte des Graphembegriffs und des frühen Phonembegriffs*. Tübingen: Niemeyer.

Li, X. and W. Shen (2013). “Joint effect of insertion of spaces and word length in saccade target selection in Chinese reading.” In: *Journal of Research in Reading* 36.1, pp. 64–77.

Li, Xingshan et al. (2015). “The role of words in Chinese reading.” In: *Oxford library of psychology. The Oxford handbook of reading*. Ed. by Alexander Pollatsek and Rebecca Treiman. Oxford: Oxford University Press, pp. 232–244.

Liu, P. and X. Li (2014). “Inserting spaces before and after words affects word processing differently: Evidence from eye movements.” In: *British Journal of Psychology* 105, pp. 57–68.

McCarthy, John J. and Alan Prince (1995). “Faithfulness and reduplicative identity.” In: *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*. Ed. by Jill Beckman, Laura Dickey, and Suzanne Urbanczyk. Amherst: GLSA, pp. 249–384.

Meng, Yuxian et al. (2019). “Is Word Segmentation Necessary for Deep Learning of ChineseRepresentations?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3242–3252.

Nespor, Marina and Irene Vogel (1986). *Prosodic Phonology*. Fordrecht: Foris.

Osaka, Naoyuki (1989). “Eye fixation and saccade during kana and kanji textreading: comparison of English and Japanese text processing.” In: *Bulletin of the Psychonomic Society* 27, pp. 548–550.

——— (1992). “Size of saccade and fixation duration of eye movements during reading: psychophysics of Japanese text processing.” In: *Journal of the Optical Society of America A* 9.1, pp. 5–13.

Packard, Jerome L. (1998). “Introduction.” In: *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*. Ed. by Jerome L. Packard. Berlin: De Gruyter, pp. 1–34.

——— (2000). *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge: Cambridge University Press.

Primus, Beatrice (2003). “Zum Silbenbegriff in der Schrift-, Laut- und Gebärdensprache – Versuch einer mediumübergreifenden Fundierung.” In: *Zeitschrift für Sprachwissenschaft* 22, pp. 3–55.

——— (2010). “Strukturelle Grundlagen des deutschen Schriftsystems.” In: *Schriftsystem und Schrifterwerb: linguistisch – didaktisch – empirisch*. Ed. by Ursula Bredel, Astrid Müller, and Gabriele Hinney. Tübingen: Niemeyer, pp. 9–45.

Prince, Alan and Paul Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Ms., Rutgers University (= Rutgers University Center for Cognitive Science Technical Report 2).

Ratcliffe, Robert R. (2001). “What do “phonemic” writing systems represent?” In: *Written Language & Literacy* 4.1, pp. 1–14.

Rogers, Henry (2005). *Writing Systems: A Linguistic Approach*. 3rd ed. Oxford: Blackwell.

Rollings, Andrew G. (2004). *The spelling patterns of English*. Muenchen: Lincom Europa.

Roubah, Aïcha and Marcus Taft (2001). “The functional role of syllabic structure in French visual word recognition.” In: *Memory & Cognition* 29, pp. 373–381.

Ryan, Des (2018). “Principles of English spelling formation.” PhD thesis. Trinity College Dublin.

Sainio, Miia et al. (2007). “The role of interword spacing in reading Japanese: An eye movement study.” In: *Vision Research* 47, pp. 2575–2584.

Schiering, René, Balthasar Bickel, and Kristine E. Hildebrandt (2010). "The prosodic word is not universal, but emergent." In: *Journal of Linguistics* 46.3, pp. 657–709.

Sproat, Richard (2010). *Language, technology and society*. Oxford: Oxford University Press.

Taylor, John R. (2015). "Introduction." In: *Oxford Handbook of the Word*. Ed. by John R. Taylor. Oxford: Oxford University Press, pp. 1–23.

Weingarten, Rüdiger (2004). "Die Silbe im Schreibprozess und im Schriftspracherwerb." In: *Schriftspracherwerb und Orthographie*. Ed. by Ursula Bredel, Gesa Siebert-Ott, and Tobias Thelen. Baltmannsweiler: Schneider, pp. 6–21.

Wurzel, Wolfgang Ulrich (2000). "Was ist ein Wort?" In: *Deutsche Grammatik in Theorie und Praxis*. Ed. by Rolf Thieroff et al. Tübingen: Niemeyer.

Zhang, S. (1985). "On some problems with Chinese phonetic writing orthography." In: *Chinese Phonetic Writing Group*, pp. 61–66.

Zifonun, Gisela, Ludger Hoffmann, and Bruno Strecker, eds. (1997). *Grammatik der deutschen Sprache*. Berlin/New York: De Gruyter.