Foto: Lisa Beller

Universität zu Köln

What is a written word? And if so, how many?

Martin Evertz-Rittich | University of Cologne

/gʁafematik/ Grapholinguistics in the 21st century | 17.06.2020

Outline

1. Defining the written word in alphabetical writing systems
2. Properties of written words
3. Correspondence to elements in spoken language
4. Typological considerations
5. Summary

Defining the written word in alphabetical writing systems

Part I

Definition by spaces

(e.g. Coulmas 1999, 550; Jacobs 2005, 22; Fuhrhop 2008, 193f.)

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

Problems:

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

Definition by spaces

(Zifonun et al. 1997, 259; my translation)

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

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

Problems:

- <you.>, <you?>, <you!>
- <Smiths’> (e.g. in the Smiths’ house), <mother-in-law>
- <“you”>, <(you)>

Towards a typographic definition: fillers and clitics

- Characters and punctation marks can be divided into two classes (Bredel 2009)
- Fillers
  - They can independently fill a segmental slot
  - Letters, numbers, apostrophes, hyphens
- Clitics
  - They need the support of a filler
  - periods, colons, semi-colons, commas, brackets, question marks, quotation marks, exclamation marks

A typographic definition

Evertz (2016a, 391-392 based on works of Bredel; my translation)

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

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!	

A typographic definition – consequences

Evertz (2016a, 391-392)

- Distinction between **graphic surface** and **graphematic word**
- Clitics are part of the graphic surface but they are not part of the graphematic word
- Fillers are part of the graphic surface **and** the graphematic word
  - That is true for **all** fillers including non-letter fillers

A typographic definition – solutions to former problems

cf. Evertz (2016a, 391-392)

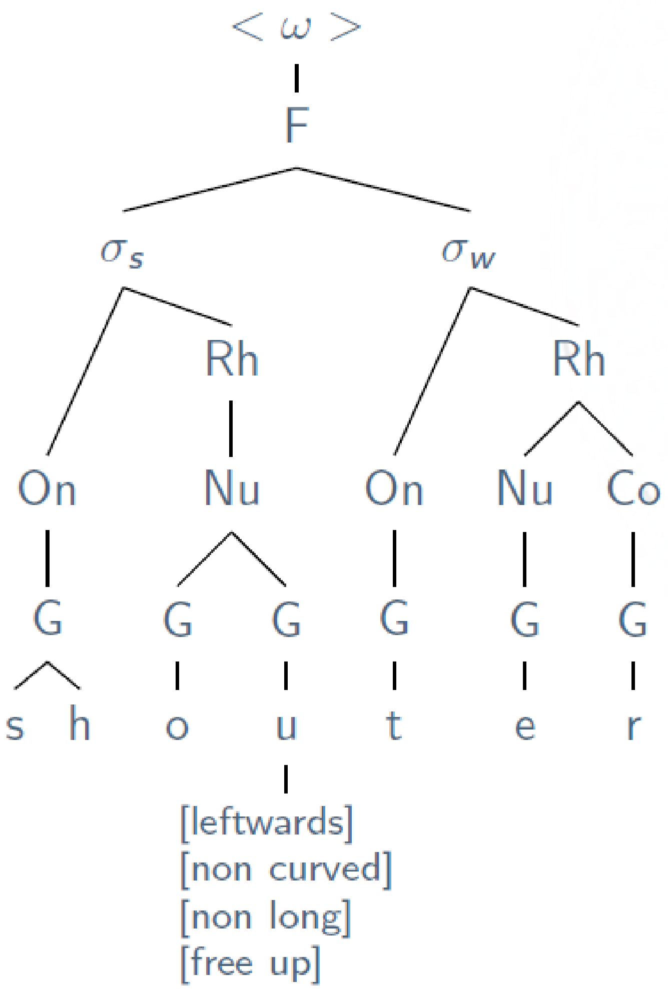
- |**you**.|, |**you**?|, |**you**!|, |“**you**”|, |(**you**)|
  - **one** graphematic word <**you**> with different graphic surfaces

- <Smiths**’**> (e.g. in the Smiths’ house), <mother**-**in**-**law>
  - Apostrophe and hyphen are part of the graphematic word
    - Apostrophe signals that some information is missing
    - Hyphen signals that the morphological processing of the word is not completed

Properties of graphematic words

Part II

Graphematic hierarchy (cf. Evertz & Primus 2013, Evertz 2018)



Word level

Foot level

Syllable level

Grapheme level

Segmental level

Feature level

- Suprasegmental units in phonology and graphematics are **hierarchically** organized
- Every nonterminal unit of the hierarchy is composed of one or more units of the immediately lower category (cf. Nespor & Vogel 1986, 7)

Graphematic hierarchy – consequences

(4) A graphematic word consists of at least one graphematic foot.

(5) A graphematic foot consists of at least one graphematic syllable.

- It follows that a graphematic word has to conform to well-formedness constraints of syllables and feet

Example: minimal weight

Evertz (2016b)

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

(6) Content words must have more than two letters.
(e.g. Cook 2004, 57)

- Explanation:
  - A content word consists of at least one graphematic foot
  - In order to constitute a monosyllabic foot, a syllable needs to have a graphematic minimal weight (it must be bimoraric)
  - Thus, a monosyllabic word needs to have a certain minimal weight

Exceptional words

- The constraints pertaining to the well-formedness of syllables and feet (5-6) are **violable**
  - Ill-formed graphematic syllables: *Mr., Mrs., vs., Dr.*
  - Ill-formed graphematic feet: *BA, MA, no.*
- Exceptions to (5-6) may be licensed through special orthographic devices like dots or all-caps

Correspondence to elements in spoken language

Part III

Correspondents of the graphematic word

Fuhrhop (2008), Fuhrhop & Peters (2013), Evertz (2016a)

- The graphematic word mainly corresponds to the morphological or syntactical word in German
- Writer‘s perspective:
  - Separate syntactic words by empty slots
  - Write morphological words without empty slots in between
- Reader‘s perspective:
  - Interpret slot-filler-sequences without spaces **morphologically**
  - Interpret slot-filler-sequences with spaces **syntactically**

wohlgeraten ‘great, outstanding‘

- no empty slots within
- one graphematic word
- one morphological word

wohl geraten ‘probably guessed‘

- empty slot between words
- two graphematic word
- syntactical phrase

English compounds

- Only little free variation
  - e.g. <secondhand>, <second-hand>, <second hand>
- Compounds are generally hyphenated or written without empty slots. Open writing is most often motivated by the avoidance of length (cf. Sanchez-Stockhammer 2018)
- Using the hyphen or writing without empty slots can help to avoid ambiguity
  - <blackbird>, <black bird>
  - <old furniture dealer>, <old furniture-dealer>, <old-furniture dealer>
- Thus, it seems that the graphematic word in English also corresponds to the syntactic and morphological word

Typological considerations

Part IV

Non-alphabetical writing systems

- The presented definition of a graphematic word seems to be useful for (most of) alphabetical writing systems
- In some writing systems, however, there are no empty slots, so the definition in (3) cannot apply
- This might be due to linguistic features of the corresponding spoken languages or because of certain features of these writing systems

Chinese writing system

cf. Chen (1996), Li et al. (2015)

- A Chinese character represents most likely a morpheme or a syllable
  - 蚯蚓 *Qiūyǐn* ‘earthworm‘: neither character represents a morpheme (Chen 1996, 46)
- Approximately 97% of words in Chinese are one or two characters in length (token frequency; Lexicon of Common Words in Contemporary Chinese Research Team, 2008)
- The majority of modern Chinese words are bi-morphemic: ca. 80% (Li 1977)
- Words are not marked by empty slots

Example sentence

Coulmas (2003, 59)

中国这几年的变化的确很大。

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‘

Linguistic features of Chinese

Hoosain (1992), Chen (1996), Packard (2000, 2015)

- Chinese almost completely lacks inflection
- Morphemes in Chinese can be *free* or *bound*
  - There are degrees of freedom
  - The status of a morpheme as free or bound can vary by context, register and dialect
- Bound morphemes may occur before or after a free morpheme
- These factors contribute to a “fluidity of word boundaries” in Chinese (Hoosain 1992, 120; Chen 1996, 46)

Historical reasons

- Classical Chinese was mostly monosyllabic and monomorphematic, thus words and characters were almost congruent (Hoosain 1992, 119; Li et al. 2015, 232)
- There was no term for a word in Chinese until the concept was imported from the West at the beginning of the twentieth century (Packard, 1998)
    - Note: 字 *zì* ‘morpheme-syllable, character‘ ≠ 词 *cí* ‘syntactic word‘ (Packard 2000)

Further reasons

Li et al. (2015, 232-233)

- The variance in word length is reduced relative to word length variability in alphabetic languages
- The number of potential sites within a character string at which word segmentation might occur is significantly reduced in Chinese
- Therefore 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

Psycholinguistic evidence

- Word spaced text (or highlighting) does not facilitate reading Chinese, but did not interfere with reading in adult readers
  (Inhoff et al. 1997; Bai et al. 2008)
- 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 & Shen, 2013; Liu & Li, 2014)

Japanese writing system

e.g. Joyce & Masuda (2018)

- There are mainly two kinds of characters in Japanese: **kana** and **kanji**
- Most kanji are associated with lexical morphemes
- Okurigana (hiragana) are used for high-frequency morphemes such as postpositions and inflectional endings
- Katakana are mainly used for non-Chinese loanwords

Japanese writing system

- Because of the different scripts within the JWS, readers may easily differentiate between **content** and **grammatical** elements (Joyce & Masuda 2016)
- Kanji are **visually salient** (Kaji et al. 2001)
- The **word-beginning** is typically occupied by a kanji (Rogers 2005, 66)
- Thus, characters, frequently appearing in the word beginning, serve as effective **segmentation cues** to signal word boundaries (Sainio et al. 2007)

Example sentence

Shibatani (1990, 129), Rogers (2005, 66)

K	hg			kk	hg	K	hg				rom	hg	
花子	は	あ	の	ビル	で	働	い	て	い	る	OL	で	す。
Hanako	wa	a	no	biru	de	hatari-	i-	te-	i-	ru	ooeru	de	su
Hanako	topic	that		building	at	work-	ing				OL	is	

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

K = kanji, hg = hiragana, kk = katakana, rom = Roman

Psycholinguistic evidence

Sainio et al. (2007)

- Japanese readers are facilitated by interword spacing when reading texts written exclusively in syllabic kana…
- …but **not** with texts that are written in the normal mixture of *kana* and *kanji*

Summary

- Chinese
  - Morphemes seem to be more **salient** than words in Chinese grammar
  - In classical Chinese, morphemes, words and characters were almost congruent
  - Thus, the morpheme/syllable is marked rather than the word
- Japanese
  - Word boundaries are **graphotactically** marked in Japanese
  - Interword separation by spaces or other punctuation marks (e.g. interpunct) are therefore unnecessary
- English/ German
  - Words are salient units in English & German grammar
  - There are no graphotactical means to indicate word boundaries

Summary

Part V

Summary

- With a **typography-based definition,** graphematic words can be defined in alphabetical writing systems
- Properties of graphematic words can be deduced from the **graphematic hierarchy**
- The graphematic word corresponds to the **morphological and syntactic word**
- Writing systems without interword spacing most likely lack spacing because of **linguistic features** or because they already have **cues to word boundaries** that make spacing unnecessary

Thank you for your attention!

Bibliography

- Bredel, Ursula (2009): Das Interpunktionssystem des Deutschen. In Angelika Linke & Helmuth Feilke, *Oberfläche und Performanz*. Tübingen, 117–135.
- Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance* 34, 1277–1287.
- Chen, May J. (1996): An overview of the characteristics of the Chinese writing system. *Asia Pacific Journal of Speech, Language and Hearing* 1(1), 43-54.
- Coulmas, Florian (1999): *The Blackwell encyclopedia of writing systems*. Oxford (UK)/ Cambridge (Mass.).
- Coulmas, Florian (2003): *Writing Systems: An Introduction to Their Linguistic Analysis*. Cambridge.
- Evertz, Martin (2016a): Graphematischer Fuß und graphematisches Wort. In Beatrice Primus, & Ulrike Domahs (eds), *Laut – Gebärde – Buchstabe*. Berlin/ New York, 377-397.
- Evertz, Martin (2016b): Minimal graphematic words in English and German: Lexical evidence for a theory of graphematic feet. *Written Language and Literacy* 19(2), 189-211.
- Evertz, Martin (2018): *Visual Prosody – The Graphematic Foot in English and German*. Berlin/ New York.
- Evertz, Martin & Beatrice Primus (2013): The Graphematic Foot in English and German. *Writing Systems Research 5(1)*, 1–23.
- Fuhrhop, Nanna & Jörg Peters (2013): *Einführung in die Phonologie und Graphematik*. Stuttgart.
- Fuhrhop, Nanna (2008): Das graphematische Wort (im Deutschen): Eine erste Annäherung. In *Zeitschrift für Sprachwissenschaft* 27, 189–228.
- Hoosain, Rumjahn (1992): Psychological reality of the word in Chinese. In HC Chen & OJL Tzeng (eds.) *Language processing in Chinese*. Amsterdam, 111-130.
- Inhoff, A., & Wu, C. (2005): Eye movements and the identification of spatially ambiguous words during Chinese sentence reading. *Memory & Cognition* 33, 1345-1356.
- Jacobs, Joachim (2005): *Spatien. Zum System der Getrennt- und Zusammenschreibung im heutigen Deutsch*. Berlin/New York.
- Joyce, Terry, & Masuda, Hisashi (2016): Just mixed up or a pretty neat idea? Some reflections on the multi-script nature of the Japanese writing system. Presentation given at '*Understanding writing systems: From core issues to implications for written language acquisition' – 10th International Workshop on Written Language and Literacy*, 12–13 May, Radboud University, Nijmegen, The Netherlands.
- Joyce, Terry, & Masuda, Hisashi (2018): Introduction to the multi-script Japanese writing system and word processing. In H. Pae, (ed.) *Writing Systems, Reading Processes, and Cross-Linguistic Influences. Reflections from the Chinese, Japanese and Korean Languages*. Amsterdam, 179-200.
- Kajii, Natsumi, Nazir, Tatjana A. & Osaka, Naoyuki (2001). Eye movement control in reading unspaced text: The case of the Japanese script. *Vision Research* 41, 2503–2510.
- Li H.T. (1977): *The History of Chinese Characters*. Taipei, Taiwan: Lian-Jian.
- Li, Xingshan, Zang, Chuanli, Liversedge, Simon P. & Pollatsek, Alexander (2015): The role of words in Chinese reading. In Alexander Pollatsek & Rebecca Treiman (Eds.), *Oxford library of psychology. The Oxford handbook of reading* (p. 232–244). Oxford University Press.
- Li, X., & Shen, W. (2013). Joint effect of insertion of spaces and word length in saccade target selection in Chinese reading. *Journal of Research in Reading* 36(S1), S64–S77.
- Liu, P., & Li, X. (2014). Inserting spaces before and after words affects word processing differently: Evidence from eye movements. *British Journal of Psychology* 105, 57–68.
- Nespor, Marina & Irene Vogel (1986): *Prosodic Phonology*. Dordrecht.
- Packard, Jerome L. (1998): Introduction. In J. L. Packard (Ed.), *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*. Berlin, 1-34.
- Packard, Jerome L. (2000). *The Morphology of Chinese: A Linguistic and Cognitive Approach.* Cambridge.
- Packard, Jerome L. (2015): Morphology: Morphemes in Chinese. In William S-Y. Wang & Chaofen Sun (eds.), *The Oxford Handbook of Chinese Linguistics*, 263-274.
- Rogers, Henry (2005): *Writing Systems: A Linguistic Approach*. Malden (MA), Oxford, Victoria (Australia): Blackwell.
- Sainio, Miia, Hyönä, Jukka, Bingushi, Kazuo& Bertram, Raymond (2007): The role of interword spacing in reading Japanese: An eye movement study. *Vision Research* 47, 2575–2584.
- Sanchez-Stockhammer, Christina (2018). *English Compounds and their Spelling* (Studies in English Language). Cambridge.
- Shibatani, Masayoshi (1990): *The languages of Japan*. Cambridge: University Press.
- Wiese, Richard (2000): *The Phonology of German*. 2nd rev. ed. Oxford, UK.
- Zifonun, Gisela/Ludger Hoffmann/Bruno Strecker, u. a. (Hg.) (1997): *Grammatik der Deutschen Sprache*. Berlin/New York.

Appendix

Towards a typographic definition: fillers and clitics

- Characters and punctation marks can be divided into two classes (Bredel 2009)
- Fillers
  - They are symmetric, i.e. to the left *and* right of a filler can be elements of the same class. Examples: <abc**-**def>, <a**b**c>
  - They can independently fill a segmental slot
  - Letters, numbers, apostrophes, hyphens
- Clitics
  - They are asymmetric. Examples: *<abc**.**def>, *<abc**!**def>
  - They need the support of a filler
  - periods, colons, semi-colons, commas, brackets, question marks, quotation marks, exclamation marks

Phonological word ≠ graphematic word

- Phonological word: Domain for phonological rules such as syllabification
    - Onset maximisation: intervocalic consonants are maximally assigned to the onsets of syllables
- Example: *Tierart* ‘animal species‘ (Wiese 2000, 65 f.)
    - [ˈtiːɐ̯.ʔaːɐ̯t] vs. *[ˈtiː.raːɐ̯t]
    - {Tier}{art}
- Thus: graphematic and phonological word do not map exactly unto each other

Morphological word?

Fuhrhop (2008, 224)

- Morphological word
  - Inflecting uniformly (Wurzel 2000, 36)
  - Constituted due to word building rules (Jacobs 2005)
- Example: *Tierart* ‘animal species‘
  - Inflecting uniformly: *Tierarten* vs. **Tierearten*
  - Constituted due to composition rules
  - Morphological word and graphematic word
- Possible exception: *Langeweile* ‘boredom‘
  - *(mit seiner) ?Langenweile* ‘with his boredom (Dativ)‘ (Wurzel 2000, 57)

Syntactic word?

Fuhrhop (2008, 193)

- Syntactic word
  - syntactically free form, commonly designated in the literature as X^0
- Example:

er	*fängt*	*mit*	*dem*	*Schreiben*	*an*
he	*starts*	*with*	*the*.DAT	*writing*	PTCL

‘he starts writing‘

- **an fängt er mit dem Schreiben*
  - The particle *an* is not a syntactic word (not permutable, part of the verb)
  - It is, however, a graphematic word

The CompSpell algorithm

Sanchez-Stockhammer (2018, 352), my emphasis

- Adjective (broken-down) **Hyphenated**
  Adverb (well-nigh)
  Verb (chain-smoke)
- Noun
  - **three or more syllables (bathing suit)** **Open**
  - two syllables
    - second constituent: up to two letters (close-up) **Hyphenated**
    - second constituent: more than two letters (coastline) **Solid**

Accuracy: 61%-80.7% depending on corpus

Thai language and writing system

Danvivathana (1981, 269), Smyth (2014, 1-2), Kasisopa et al. (2016, 72)

- Language
    - No noun or verb inflections
    - Tonal language
    - Average word-length ca. 3 to 4 syllables
        - Native words are mostly monosyllabic
        - Borrowings most often polysyllabic
    - many compound words
- Writing system
    - Alphabetic writing system
    - no empty slots between words
    - when empty slots are used, they serve as punctuation markers, instead of commas or full stops
        - empty slots are normally used at the end of a phrase, clause or a sentence

Cues to syllables in Thai writing system

Slayden (2010)

- Following vowels start a syllable: <เ, แ, โ, ใ, ไ>
  - <ใ> and <ไ> start an open syllable
- <ะ>, <อ์ > and <อำ> end a syllable (exceptions exist)
- <อั> and <อ็> do not appear over a syllable final consonant
- Two consonants may form an initial cluster; a tone mark, if any, will appear on the second consonant of such a cluster

Psycholinguistics of Thai reading

- Adding spaces between words facilitates reading rates
  (Kohsom & Gobet,1997)
- Word-initial and word-final position-specific frequency of consonants may be used as cues to word boundaries
  (Reilly et al. 2005, Kasisopa et al. 2016)
- Thai readers employ a flexible targeting system (for eye fixation) that makes opportunistic use of available statistical cues to the location of words and their centers
  (Kasisopa et al. 2016, 80)
  - The position-specific frequencies of word-initial and word-final characters assist in directing Thai readers to an optimal viewing position just left of word center

Summary: Thai

- The native lexicon of Thai is mainly composed of monosyllabic words
- Thai is an analytic language
- There are robust cues to identify syllable boundaries in the Thai writing system
- Thus, there was (and is) no need to mark words by empty slots

Bibliography (Appendix)

- Danvivathana, Nantana (1981): *The Thai Writing System*. Dissertation University of Edinburgh.
- Kasisopa, Benjawan, Reilly, Ronan G., Luksaneeyanawin, Sudaporn & Burnham, Denis (2016): Eye movements while reading an unspaced writing system: The case of Thai. *Vision Research* 86, 71–80.
- Kohsom, Chananda & Gobet, Fernand (1997): Adding Spaces to Thai and English: Effects on Reading. *Proceedings of the Cognitive Science Society*, 19, 388–393.
- Reilly, Ronan G., Radach, Ralph, Corbic, D., & Luksaneeyanawin, Sudaporn (2005): Comparing reading in English and Thai: The role of spatial word unit segmentation in distributed processing and eye movement control. Paper presented to ECEM 13. University of Bern, 13–18 August, 2005.
- Slayden, Glenn (2010): How do I recognize where Thai words begin and end? http://www.thai-language.com/ref/breaking-words (retrieved 17.06.2020).
- Smyth, David (2014): *Thai: An Essential Grammar*. London.