15th AWLL international workshop on writing systems and literacy

25.10.2025

Jonas Romstadt (jonas.romstadt@uol.de)

BREAKING WRITTEN WORDS

Institute for Computational Linguistics *Antonio Zampolli*, Pisa, Italy

“[This] constitutes **a rather marginal aspect** of our orthography. In writing, it can always be avoided, and many word processing programs now perform it entirely automatically for the writer.”

Eisenberg ([5]2020: 333)

[redacted]ng

Deutscher Aufsatz

Vorschlag: 3
Vorzensur: 3
Schriftl. Jahresleistg.: 2, 2, 3, 3

Deutscher Reifeprüfungsaufsatz

Zu welchen Gedanken werden Sie durch den in letzter Zeit häufig zitierten Wahlspruch des Papstes angeregt: Gerechtigkeit schafft Frieden? Denken Sie dabei an das Zusammenleben der einzelnen Menschen und der Völker!

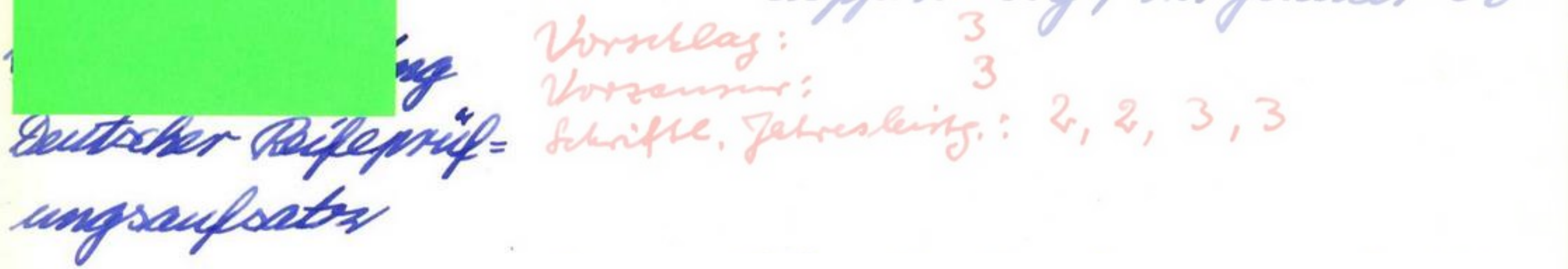
A Ein Wahlspruch, der als Leitge=

Deutscher Reifeprüf=
ungsaufsatz

Vorschlag: 3
Vorzensur: 3
Schriftl. Jahresleistg.: 2, 2, 3, 3

To what reflections are you prompted by the Pope's recent and frequently quoted maxim: "Justice creates peace"? In your response, consider both the coexistence of individuals and that of nations.

A Ein Wahlspruch, der als Leitge=



German matriculation essay

Deutscher Reifeprü≈ fungsaufsatz

To what reflections are you prompted by the Pope's recent and frequently quoted maxim: "Justice creates peace"? In your response, consider both the coexistence of individuals and that of nations.

The graphematic syllable boundary

- the **graphematic syllable** is defined by its nucleus, which contains a compact letter (in the present case ‹u› or ‹ü›)
- for hyphenation, there are two potential division points, both of which can be justified from a purely graphematic perspective
- variant 1 “is a possible purely graphematic syllabification which is additional justified from a functional perspective because it codes the phonological syllabification” (Schmidt 2014: 270); variant 2 “is a possible possible purely graphematic syllabification too, which is also justified from a functional perspective because it codes the morphological segmentation of the word.” (ibid.)

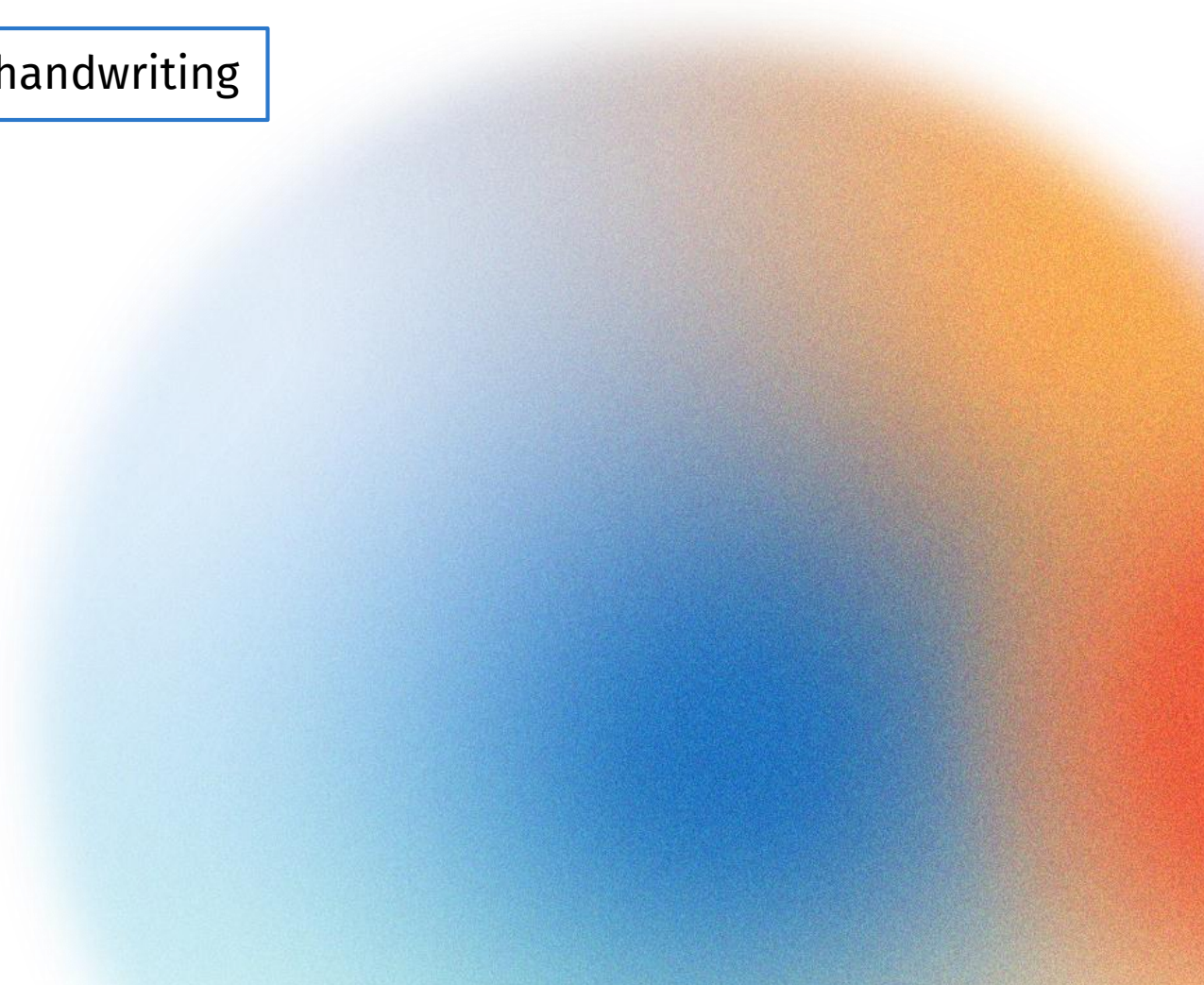
Schmidt (2014), Fuhrhop/Schmidt (2014)

The graphematic syllable boundary

- the graphematic syllable boundary is therefore **relational**: “no syllabification should be privileged. In addition, and this is crucial, from a functional perspective both syllabifications are well justified.” (Schmidt 2014: 272)
- this can be systematically substantiated: there is “no necessity to define the syllable boundary precisely, and in particular, no reason to define the syllable by means of the syllable boundary. Syllables are defined by their nucleus.” (Fuhrhop/Schmidt 2014: 541)
- we know from psycholinguistic studies that readers make use of such boundaries when processing written texts; see, among others, Bredel/Noack/Plag (2013) and Geilfuß-Wolfgang (2007) for German, as well as Diependaele/Grainger/Sandra (2011) and Grainger/Ziegler (2011) for English

Outline

- From the graphematic syllable boundary to line-end hyphenation
- Text mode vs. List mode
- Breaking written words in handwriting
- Conclusions



Line-end hyphenation

Graphematic syllables and line division

- in principle, line-end hyphenation is bound to graphematic syllables: “Line-end division indicates possible graphematic syllables” (Fuhrhop/Schmidt 2014: 540–541)
- the division itself reflects the relational nature of the syllable boundary in turn

“Between vowel graphemes there is a division point. If consonant graphemes are present, the division occurs before the last one.” (Eisenberg [5]2020a: 335)

“In forms with prefixes, verb particles, and compounds, syllable division takes place at the morphological boundary.” (ibid.)

(1) Prü-fungen
(1952_DEU_K2_03_M_08P)

(2) Reifeprüf-ungsaufsatz
(1958_DE_K1_14_M_08P)

(3) Vergessen-heit
(2018_GE_LK2_15_W_05P)

“We speak of a rule of division rather than a regularity, because it must remain open how far it actually extends and on what basis it is founded. It is a kind of general guideline [...].” (Eisenberg [5]2020: 335)

Interlude: Text mode and List mode

- written texts differ according to whether they are **continuous** or not
- for continuous texts, it holds that: “If there were no page end, the elements could appear in an almost endlessly long sequence without anything changing in meaning or grammaticality.” (Reißig 2015: 33)

Text-modal texts (continous)

A maxim intended to serve as a guiding principle for a person's actions points both to the past and the future; to the past in the sense that the maxim is drawn from the experience one has accumulated, and no one can suddenly completely deny their own nature.

List-modal texts (non-continous)

To-Do-List
prepare presentation for the conference in Pisa
go shopping
feed the cat

Reißig (2015)

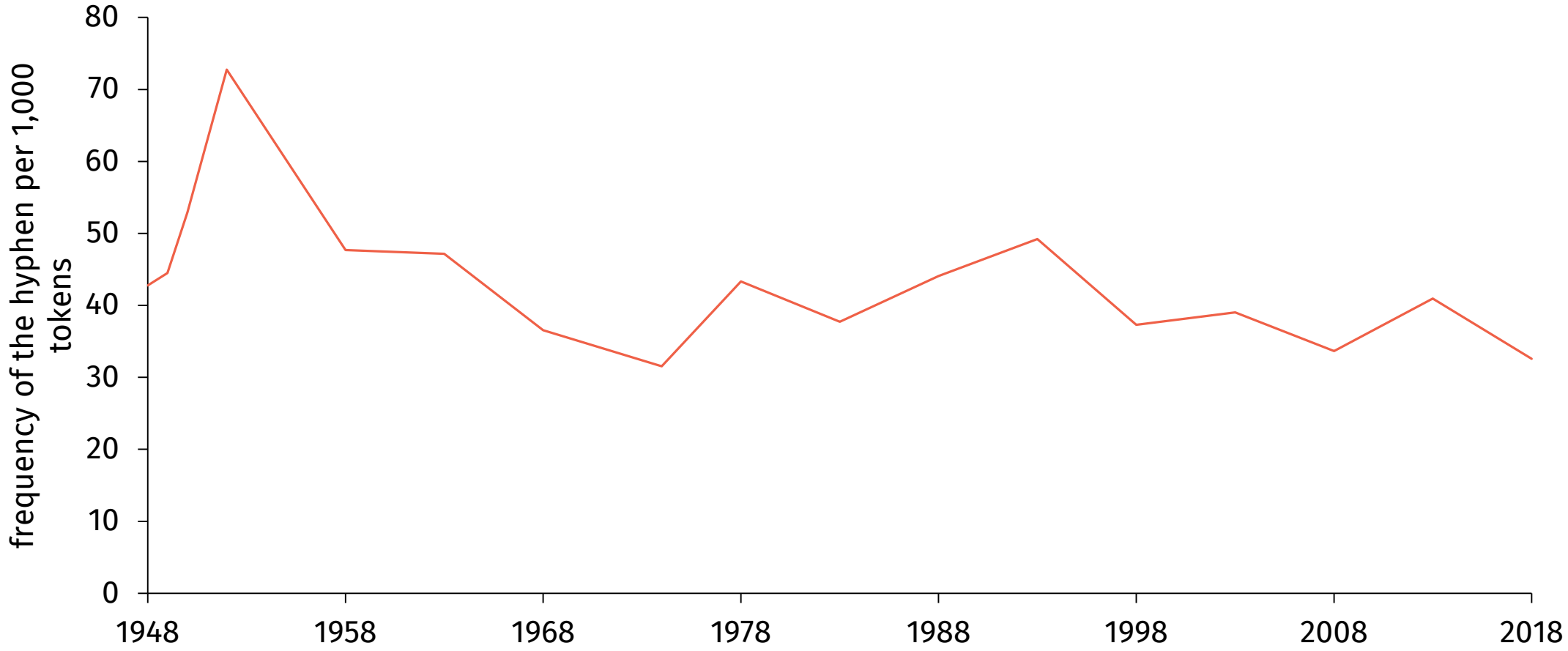
Usage-based graphematics

- “professionally” produced texts (such as newspapers) have usually gone through a final proofreading process
- texts produced by non-professionals may have been created using the help of spell checkers or AI text production
- texts written by learners are interesting, but they represent different stages in the learning process (cf. Müller 2007)
- additionally, in most written corpora, line-end hyphenations are normalized, and thus no longer detectable

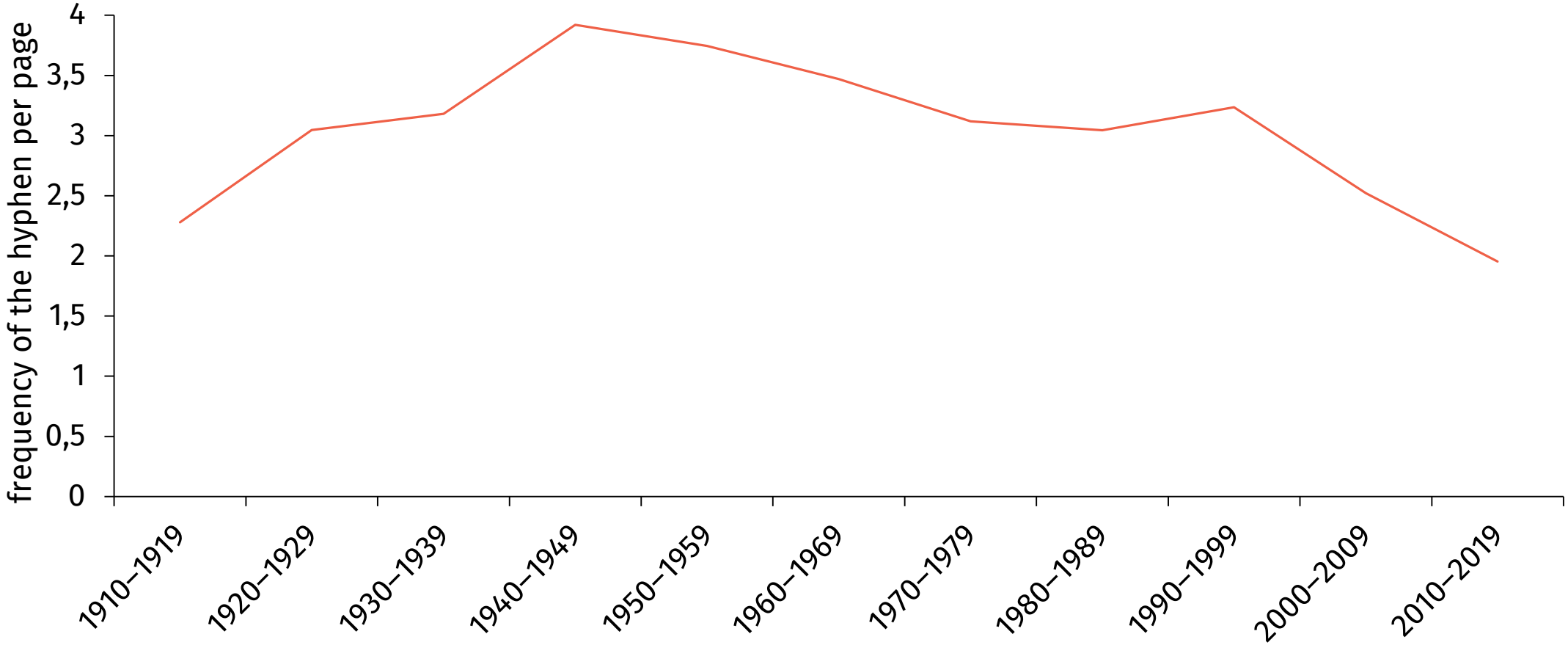
The examination of a-level exams (“Abiturklausuren”) avoids these problems: they are handwritten under comparable conditions for more than 100 years …

… by writers who try to produce a text as correct as possible (cf. Sebba 2007) and who are at the end of their institutionally supervised learning process.

Global frequency of the hyphen



Frequency of the hyphen per page



Separated words

	all graphematic words *n* = 2,035,073	line-separated graphematic words *n* = 48,769	unseparated graphematic words *n* = 1,986,324	104 most frequent line-separated words
mean	5.68 letters	11.92 letters	5.52 letters	11.84 letters
deviation	3.48 letters	4.10 letters	3.31 letters	3.83 letters

Separated words

	occurrences separated	total occurrences	relative
Menschen	185	4611	4.0 %
Gesellschaft	163	1510	10.8 %
Deutschland	156	2029	7.7 %
werden	105	7902	1.3 %
gegenüber	96	942	10.2 %
Entwicklung	88	893	9.9 %
deutlich	79	1952	4.0 %
Photosyntheseleistung	78	245	31.8 %
Sauerstoff	78	1219	6.4 %
Nationalsozialismus	76	298	25.5 %
unterschiedlichen	76	421	18.1 %
...	...	...	...

Separated words

Reifeprüfungsaufsatz (*matriculation essay*)		***Prüfungsaufsatz*** (*examination essay*)	
Reife-prüfungsaufsatz	5		
Reifeprü-fungsaufsatz	5		
Reifeprüf-ungsaufsatz	1	Prüf-ungsaufsatz	1
		Prüfung-saufsatz	1
Reifeprüfungs-aufsatz	33	Püfungs-aufsatz	58
Reifeprüfungsauf-satz	4	Prüfungsauf-satz	10

Sonneneinstrahlung (*solar radiation*)		***Photosyntheseleistung*** (*photosynthetic performance*)	
Sonn-eneinstrahlung	1	Pho-tosyntheseleistung	2
Snnen-einstrahlung	15	Photo-syntheselistung	33
Sonnenein-strahlung	10	Photosyn-theseleistung	6
Sonneneineinstrah-lung	5	Photosynthese-leistung	34
		Photosyntheselei-stung	2
		Photosyntheseleis-tung	1

Separated words

Standesunterschiede (*social class differences*)	
Stan-desunterschiede	9
Standes-unterschiede	20
Standesun-terschiede	3
Standesunter-schiede	6
Standesunterschie-de	2

Menschen (*people*)	
Men-schen	185

Gesellschaft (*society*)	
Ge-sellschaft	43
Gesell-schaft	120

Deutschland (*Germany*)	
Deut-schland	4
Deutsch-land	152

werden (*become*)	
wer-den	105

gegenüber (*compared to*)	
ge-genüber	23
gegen-über	73

(Intermediate) Conclusion

- the most frequent placements of the hyphen occur at points where syllable and morpheme boundaries coincide
- this is also sensible, particularly from the reader's perspective, because the relational role of the graphematic syllable boundary is reflected
- at the same time, handwritten texts exhibit variation in hyphen placement
- this variation depends on the spatial conditions in which the writing process takes place
- simultaneously, there seems to be a global preference for morphological boundaries (coda maximization)
- this has been especially observed in morphologically complex units—where the need for morphological 'clarification' for readers may be particularly high …
- … especially since hyphenation makes a division point visually *especially* salient

Bibliography (1/2)

Augst, Gerhard (1997): Die Worttrennung. In: Augst, Gerhard/Blüml, Karl/Nerius, Dieter/Sitta, Horst (ed.): *Zur Neuregelung der deutschen Orthographie. Begründung und Kritik*. Tübingen: Niemeyer, 259–268. https://doi.org/10.1515/9783110927993.259

Bredel, Ursula/Noack, Christina/Plag, Ingo (2013): Morphologie lesen. Stammkonstanzschreibung und Leseverstehen bei starken und schwachen Lesern. In: Neef, Martin/Scherer, Carmen (ed.): *Die Schnittstelle von Morphologie und geschriebener Sprache*. Berlin: De Gruyter, 211–249. https://doi.org/10.1515/9783110336665.211

Diependaele, Kevin/Grainger, Jonathan/Sandra, Dominiek (2011): Derivational morphology and skilled reading: An empirical overview. In: Soivey, Michael/McRae, Ken/Joanisse, Marc (ed.): *The Cambridge handbook of psycholinguistics*. Cambridge: CUP, 311–332. https://doi.org/10.1017/CBO9781139029377.021

Eisenberg, Peter (2020): *Das Wort. Grundriss der deutschen Grammatik*. 5th edition. Stuttgart: Metzler. https://doi.org/10.1007/978-3-476-05096-0

Geilfuß-Wolfgang, Jochen (2007): *Worttrennung am Zeilenende. Über die deutschen Worttrennungsregeln, ihr Erlernen in der Grundschule und das Lesen getrennter Wörter*. Tübingen: Niemeyer.

Grainger, Jonathan/Ziegler, Johannes C. (2011): A dual-route approach to orthographic processing. *Frontiers in Psychology 2* (54), 1–13. https://doi.org/10.3389/fpsyg.2011.00054

Bibliography (2/2)

Müller, Hans-Georg (2007): *Zum „Komma nach Gefühl“. Implizite und explizite Kommakompetenz von Berliner Schülerinnen und Schülern im Vergleich*. Frankfurt am Main: Lang.

Schmidt, Karsten (2014): Morphophonographic regularities in German: the graphematic syllable boundary. A non-linear graphematic approach. *Written Language & Literacy 17* (2), 253–281. https://doi.org/10.1075/wll.17.2.04sch

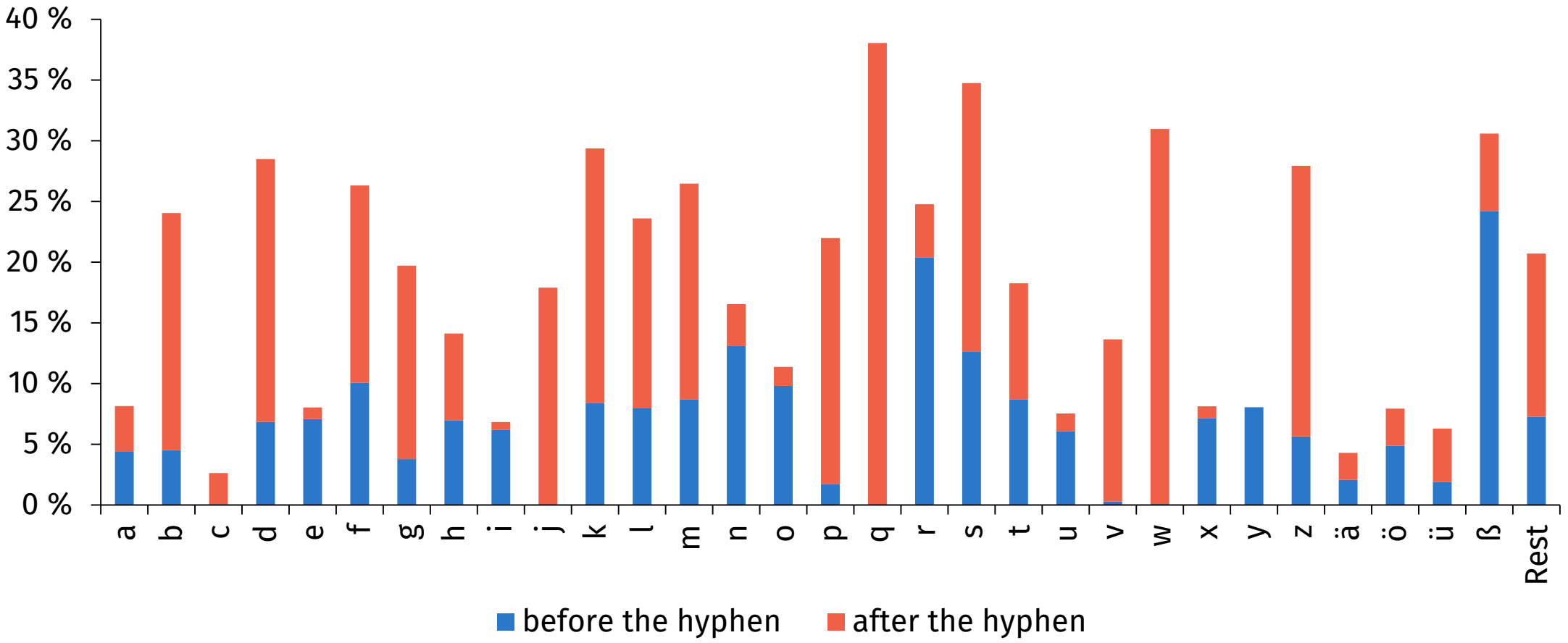
Sebba, Mark (2007): *Spelling and society: the culture and politics of orthography around the world*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511486739

Reißig, Tilo (2015): *Typographie und Grammatik. Untersuchungen zum Verhältnis von Syntax und Raum*. Tübingen: Stauffenburg.

Romstadt, Jonas/Strombach, Theresa/Berg, Kristian (2024): GraphVar – Ein Korpus für graphematische Variation (und mehr). In: Krome, Sabine/Habermann, Mechthild/Lobin, Henning/Wöllstein, Angelika (ed): *Orthographie in Wissenschaft und Gesellschaft. Schriftsystem – Norm – Schreibgebrauch*. Berlin/New York: de Gruyter, 425–435. https://doi.org/10.1515/9783111389219-024

Fuhrhop, Nanna/Schmidt, Karsten: Die zunehmende Profilierung der Schreibsilbe in der Geschichte des Deutschen. *Beiträge zur Geschichte der deutschen Sprache und Literatur 136* (4), 538–568. https://doi.org/10.1515/bgsl-2014-0047

What follows? Letters before and after the hyphen



Line-end hyphenation in standard orthography

§ 84

Multisyllabic words may be divided at the end of a line. In doing so, the boundaries of the syllables—into which the written word can be segmented when read aloud slowly—genrellay coincide with the division points.

https://grammis.ids-mannheim.de/rechtschreibung/6214

Line-end hyphenation in Optimality Theory

- the crucial constraints, which are moreover satisfied by all optimal candidates, are for German:

IDENT-TRS	VOK	*KOMPLEX-K	VERANKER-PW
graphematic „faithfulness“	at least one vowel per segment	no sequence of vowel graphemes	potential word onsets are potential division points

Moreover, Geilfuß-Wolfgang (2007) formulates six additional constraints of subordinate relevance, which are not considered here.

- “First, division points arise at prefix and compound boundaries. The remaining parts (or words that are neither prefixed nor compounded), that is, ‘simple words,’ are then segmented according to ‘syllable rules.’ A word therefore has as many potential division points as it has syllable junctures.” (Augst 1997: 260)