ZW JW Zeitschrift für Wortbildung
Journal of Word Formation

2024, 8(2), 7–27
DOI: 10.21248/zwjw.2024.2.108

Katrin Menzel

Initialisms in Scientific Writing in the 19th and Early 20th Centuries

Abstract: This paper focusses on the role of initialisms in scientific English articles in the *Royal Society Corpus* (*RSC*) from the 19th and early 20th centuries. The evolving role of scientific initialisms in English academic writing is shown here for the first time in a systematic, corpus-based analysis. The paper combines frequency data of initialisms with results from topic modelling to analyse the evolution of the topics of the texts in which initialisms are found. Additionally, it presents an analysis of information-theoretic surprisal values of initialisms in three time spans between 1830 and 1919 to measure the (un)predictability of the initialisms in their textual contexts. The results of the analysis show that the overall frequency and diversity of initialisms for scientific concepts has risen considerably between 1830 and 1919 in the context of the ongoing specialisation of the sciences. Particularly from the 1860s onwards, scientific initialisms increasingly became common shortcuts for multi-word units with wordhood and term status in a variety of disciplines of the natural sciences. The surprisal values of scientific initialisms have decreased over time as such forms more regularly occurred in conventionalised textual contexts and fixed expressions in scientific articles published by the Royal Society. Overall, the analysis confirms that key developments towards the conventionalisation of scientific initialisms as term formation patterns took place especially in the transitional period from Late Modern to Present-Day English, i.e. in the last decades of the 19th and the first decades of the 20th century.

Keywords: multi-word units, specialised languages, initialisms, diachronic word-formation, Late Modern English

1. Introduction

This paper focusses on the role of initialisms (e.g., *DRG* for ‘dorsal root ganglion’, cf. definition of initialisms below) in scientific English articles in the *Royal Society Corpus* (*RSC*, Fischer et al. 2020; Kermes et al. 2016).[1] The research questions addressed in this paper are: in which particular time span of the Late Modern English (LModE) period does the use of initialisms for scientific terms become a common strategy to shorten the growing number of multi-word terms in the natural sciences?; which scientific topics and

[1] This paper is based on research conducted in a project funded by the Deutsche Forschungsgemeinschaft, Project-ID 232722074, SFB 1102.

disciplines covered in the *RSC* were the most productive ones with regard to the use of initialisms in the analysed time period?; and have initialisms become highly predictable and conventionalised lexical items in fixed textual contexts over time? The paper aims to show that routines of shortenings for multi-word scientific term patterns have increasingly developed from the mid-19th century onwards. The evolving role of scientific initialisms in English academic writing in the 19th and early 20th centuries will be shown in a systematic, corpus-based analysis. Frequency data of initialisms and results from topic modelling will be used to analyse the evolution of the topics of the texts in which initialisms are found. Furthermore, an analysis of information-theoretic surprisal values of initialisms will be presented for three time spans between 1830 and 1919 to measure the (un)predictability of the initialisms in their textual contexts. The findings indicate a significant increase in both the frequency and variety of initialisms for scientific concepts between 1830 and 1919. Especially from the 1860s onwards, scientific initialisms increasingly became common shortcuts for multi-word units with wordhood and term status across various natural science disciplines. The surprisal values of scientific initialisms have decreased over time. This can be explained by the fact that such forms started to occur more regularly in conventionalised textual contexts and fixed expressions in scientific articles published by the Royal Society. Overall, the analysis of the *RSC* texts confirms the expectation that key developments towards the conventionalisation of scientific initialisms as term formation patterns occurred in the transitional period from Late Modern to Present-Day English (PDE). These key changes took place particularly in the last decades of the 19th century and the first decades of the 20th century.

Initialisms can be defined as combinations of initial letters of multi-word units (MWU) in condensed word-like units.[2] An example of an initialism for a scientific term is the

[2] Existing theoretical descriptions and typologies of shortening devices apply labels such as initialisms, alphabetisms, letter-words, abbreviations, alphabetic shortenings, acronyms etc. partly in slightly different ways (cf., for instance, Baum 1955; Heller & Macris 1968; Marchand 1969; Algeo 1975; Kreidler 1979; Bauer 1983; González & Cannon 1994; Fandrych 2008; Mattiello 2013). At first glance, initialisms appear to be a closed and easily identifiable category, but the real picture is much more complex and not necessarily homogeneous. Some forms depend less on the existence of a written tradition and more on the initial sounds as components. Initialisms are often grouped together with other types of shortening processes such as abbreviations, blends or multiple clippings. *Radar* ('radio detection and ranging'), for instance, takes one syllable and three initials. So-called opening letter initialisms and syllabic initialisms, e.g., *HeLa* for a cell line named after Henrietta Lacks, or *modem* ('modulator-demodulator') (cf. Bloom 2000; Hock 2021: 457), are out of the scope of this paper. The great variety of irregularly formed or hybrid forms of letter sequences that has

above-mentioned *DRG*, in which each letter corresponds to the first letter of a part from the full form 'dorsal root ganglion'. Another example is *GUT* for 'grand unified theory'. Letter-by-letter initialisms are pronounced as series of individual letter names, and acronymic initialisms as words.[3] In a broader sense, initialisms may also consist of initial letters of several lexical morphemes from multimorphematic individual words in MWU that contain more than one meaningful part. If we use such a broader definition, we may include shortenings of scientific terms that contain closed compounds (e.g., *HRP* for 'horseradish peroxidase') or neoclassical combining forms (e.g., *EMF* for 'electromotive force' or *PTFE* for 'polytetrafluoroethylene').[4]

English word-formation patterns that are productive in scientific or technical fields as well as their historical background have often been excluded from English morphological studies. The theoretical morphological literature has often treated initialisms as peripheral, marginal, or extra-grammatical word-formation patterns (Mattiello 2013). Moreover, initialisms have rarely been examined on the basis of specialised and diachronic corpus data. This paper addresses this research gap and investigates abbreviated forms in scientific writing as an increasingly regular process stipulated by changes in needs for communication due to language-internal and language-external developments. It will be shown how and when initialisms for scientific multi-word terms have become a productive word-formation pattern in specialised texts from the 19th and 20th century and spread to new contexts and usages.

Reductions of multi-word units to their initial letters as submorphemic elements (Fandrych 2004: 18) are common shortening strategies in communities that use alphabetical writing systems. Letter-by-letter initialisms often consist of three letters which makes them rather minimalistic signs. Initialisms are typically composed of capital letters (with optional periods). Initialisms are productive in specialised registers such as political, administrative, military, and business languages, and function as insiders' code words

developed in PDE makes a clear-cut definition of initialisms as the basis of an empirical approach challenging.

[3] Some initialisms have several possible pronunciations. Others neither have a pure letter-by-letter pronunciation nor a pronunciation that accurately reflects their spelling, e.g., *JPEG* pronounced as /ˈdʒeɪ ˌpeg/, *CABG* ('coronary artery bypass graft') pronounced as 'cabbage', or *CESR* ('Cornell Electron Storage Ring'), pronounced 'Caesar'.

[4] Cf. also Cannon (1989:108), who permits combining forms as constituents for forming initialisms.

giving shorter labels and an intended flavour of familiarity to concepts that already have multi-word designations (Mattiello 2013: 66). Initialisms also play an important and apparently still increasing role in scientific writing as shortening devices for MWU (Mattiello 2012; Barnett & Doubleday 2020). However, they may be perceived as in-group jargon as they are semantically less explicit than their full forms.[5]

English initialisms, especially acronymic ones, are often associated with the second half of the 20th century onwards (Fandrych 2008) and with specific communication tools and means with length constraints. Dietz (2015: 1915) describes the use of initialisms ("letter acronyms" and "word acronyms" in his terminology) as one of the few word-formation pattern "innovations" during LModE and PDE. In this article, it will be shown that the use of letter-by-letter initialisms itself was not entirely new from LModE onwards, but that their systematic use in scientific English was indeed innovative in the transitional period from LModE to PDE. Before the second half of the 19th century, scientific multi-word terms and, thus, initialisms for them were rare. Some early initialisms in English were used for non-technical terms, most importantly for multi-word proper nouns for institutions and professional and honorary designations. If we look at the first academic articles published in English at the end of the Early Modern English (EModE) period, i.e. in the second half of the 17th century, we already find some examples of initialisms. For instance, in the *Philosophical Transactions of the Royal Society of London* from 1676, the president of the Royal Society was referred to with *P.R.S.*[6] Such early initialisms from the transitional period from EModE to LModE function less word-like than later ones, as these abbreviated forms were typically not yet integrated into sentence structures and running texts.[7] They can be found in text elements such as headlines or by-lines of the articles together with other abbreviations providing information on the identity of authors. Abbreviations in

[5] Recent publications have criticised the consequences of an "exploding" (mis)use of initialisms in PDE scientific texts. Initialisms – even those that are used in highly specialised scientific fields – can be ambiguous, as the letter combinations can be used with a great variety of meanings (e.g., 16 English multi-word names of clinical trials were identified, whose short form is spelt *HEART*, cf. Fred & Cheng 2003). Moreover, some recent acronymic initialisms in scientific texts have been described as having been coined like word-formation products in marketing language (Berkwits 2000). This has also led to more mixtures of initial and non-initial letters (Tay 2020).

[6] https://royalsocietypublishing.org/doi/epdf/10.1098/rstl.1676.0043 (accessed 26 February 2024).

[7] Nevertheless, some early initialisms already underwent further word-formation processes such as conversion, derivation, or compounding so that they developed a lexeme-like behaviour and did not remain pure abbreviations (e.g., *K.C.B.-ship* and *K.C.B.'d* in 19th-century texts [*OED* s.vv. *K.C.B.-ship*, *K.C.B.*]).

compressed, heavily nominal structures can also be found in LModE 'headlinese' in other registers, for instance, in the *Old Bailey Proceedings* front matter (Hitchcock et al. 1674–1913). Initialisms that give information on professional titles of people play the most important role in the Old Bailey texts from the LModE period (e.g., *D.C.L.* for 'Doctor of Civil Law' in texts throughout the 19th century). Some initialisms for other types of multi-word proper nouns can also be found in the Old Bailey texts towards the end of the LModE period (e.g., *G.P.O.* for 'General Post Office' or *G.E.R.* for 'Great Eastern Railway').[8] At the beginning of the 20th century, there were already larger abbreviation dictionaries like Rogers's (1913) with initialisms and other shortenings from different fields.

It has been shown that the 20th century was characterised by an increasing use of multi-word terms and longer and more complex noun phrases in scientific writing (Mattiello 2012; Biber & Gray 2016). The foundations for this trend were probably laid in the 19th century when important changes in the scientific world took place that led to a rapid specialisation of scientific disciplines and journals. LModE scientific writing is already characterised by an ongoing densification of noun phrases and a growing use of multi-word terms containing proper and common nouns. The purpose of the growing number and variety of shortening devices in LModE scientific articles was to reduce the word count of texts, to save time for the authors, and to make the coding more efficient for a specialised community of readers, e.g., by reducing the wordiness of complex noun phrases and by avoiding the frequent repetition of full forms of multi-word terms with initialisms in a one-word format that achieve higher syntactic flexibility.

The analysis in this paper shows the development of scientific initialisms in the transitional period from LModE to PDE that have increasingly become shortcuts with wordhood status taking over the function of specialised vocabulary items consisting of multi-word units. Section 2 presents a diachronic case study on initialisms in the *RSC* data. The analysis focuses particularly on three 30-year periods between 1830 and 1919. In this time span, scientific initialisms were certainly not yet as frequent and diverse as in more contemporary texts. However, it is in this time period that we expect to observe crucial developments that have paved the way for the conventionalisation of initialisms for

[8] Most capital letter sequences in the Old Bailey Corpus refer to the initials of people marked on objects (e.g., "This shirt marked E.J. is my mate's, Edward Jackson's.")

scientific concepts in academic articles. After an overview on the data and methods, the development of scientific initialisms is illustrated with frequency data, an analysis of the evolution of the topics of the texts in which initialisms occur and the surprisal ranges of initialisms in the three analysed time spans. Section 3 summarises the conclusions drawn from these results.

2. Diachronic Analysis of Initialisms in Scientific Journal Articles

2.1 Data and Method

In order to gain insights into the usage of initialisms in scientific journal articles from the transitional period from LModE to PDE, the *RSC* V6.0.1 and V6.0.4 are used. The *RSC* is a large diachronic and specialised corpus of scientific English with digitised journal articles from 1665 to 1996. It contains, for instance, the *Philosophical Transactions* (*Phil. Trans.*) and the *Proceedings* (*Proc.*) of the Royal Society of London and their more specialised successor journals *Phil. Trans. A* and *B* (since 1887) and *Proc. A* and *B* (since 1905). While the early journals used to cover all major scientific disciplines of the time, the *Phil. Trans.* and *Proc. A* series published from the end of the LModE period onwards are dedicated to the mathematical and physical sciences; the *B* series cover the biological sciences. The *RSC* is a unique dataset due to its large time span and the high number of complete, professionally published and peer-reviewed science texts from many different authors. One of the advantages of the *RSC* is that it is much larger than other diachronic corpora with scientific texts, for instance, the science section in the *Representative Corpus of Historical English Registers* (ARCHER 3.2). It is also much larger than the *Coruña Corpus of English Scientific Writing* (Crespo & Moskowich 2015) that contains around 200,000 words per century and discipline from various scientific text types published during the LModE period. The *RSC* has been enriched with fine-grained annotations, e.g., for lemmas, parts of speech and metadata that provide users with contextual information.

The full corpus version V6.0.1 contains ca. 48,000 texts and ca. 296,000,000 tokens. V6.0.4 is a subcorpus from the *RSC* with all texts until 1920 (ca. 17,500 texts, 78,600,000

tokens). [9] The *RSC* V6.0.1 is annotated with surprisal values serving as an information-theoretic measure of the (un)predictability of each token in its textual context (Degaetano-Ortlieb & Teich 2022). Surprisal (S) has been calculated as the negative log (base 2) probability of each token (t) given its preceding context of three tokens measured in bits of information as in the following equation: $S(t_i) = -\log_2 p(t_i|(t_{i-1}\ t_{i-2}\ t_{i-3}))$. The texts in V6.0.4 are annotated with primary and secondary topics derived from probabilistic topic modelling that serve as indicators of the scientific topics and disciplines of the texts (Fankhauser, Knappen & Teich 2016; Menzel, Knappen & Teich 2021). The data used for the case study on initialisms in this paper are three 30-year time slices from 1830 to 1919 from *RSC* V6.0.1 and V6.0.4 (Tab. 1). They represent different fields of the natural sciences and increasing proportions of more specialised mathematical, physical, and biological science texts.

Tab. 1: Size of 30-year time slices in *RSC* V6.0.1 and V6.0.4 between 1830 and 1919

Time period	Texts	Tokens
1830–1859	2,294	9,251,482
1860–1889	4,117	22,160,285
1890–1919	4,696	29,496,383

Initialisms are identified via CQP queries (Corpus Query Processor, cf. Evert 2005) and extracted from *RSC* V6.0.1 in the three respective time slices between 1830 and 1919. Their frequencies are normalised. The full forms of the initialisms are determined manually.[10] In this paper, I am particularly interested in initialisms for scientific and technical concepts as an innovation for forming specialised vocabulary items. In research on specialised languages, there is no unanimous opinion on how to separate scientific concepts clearly

[9] All *RSC* texts from the EModE and LModE period until 1920 have been made available for free download and online query in a CQPweb (cf. Hardie 2012) interface from the CLARIN-D centre at Saarland University under a persistent identifier. The full version is available onsite to researchers and students at Saarland University.

[10] The full forms are identified manually via searches in the *RSC* on the basis of the textual contexts of the initialisms, in other dictionaries such as the *OED*, or in other relevant sources such as historical or modern abbreviations dictionaries. Searches in the texts often help to identify the full form of scientific initialisms, but in many articles, they are not written out in full if the author assumes that they are known by the readership. Especially initialisms that are shortened forms of honorific titles are almost never written out in full. Geographical initialisms and many institutional initialisms also seem to have fallen under the assumed common knowledge of the readership and are rarely defined in the documents.

from other types of professional communication and from the general lexicon. Moreover, from a diachronic perspective, there might be slightly different opinions on the classification of some forms, e.g., whether *N.N.E.* 'north-north-east' should be regarded as a specialised lexical item and whether its status has changed over time. Here, initialisms for such terms that might have undergone a certain degree of determinologisation over time, i.e. a movement from specialised to general language, are not excluded from the results. However, the following categories of initialisms are excluded as I want to keep them separate from those that designate scientific concepts: initialisms denoting professional titles, ranks, and memberships of people (e.g., *M.R.C.S.* for 'Member of the Royal College of Surgeons'), personal names (*H.D.D.* for 'Henry Drysdale Dakin'), geographical entities (*U.S.A.* for 'United States of America'), publications or collections (*R.F.F.* for 'Records of Family Faculties'), and names of projects, institutions, or societies (*R.A.S.* for 'Royal Astronomical Society') if they are not used as parts of terms (e.g., *B.T.U.* for 'Board of Trade unit' or *B.O.T.* ['Board of Trade'] *cells*).[11]

For practical reasons, the analysis focuses on initialisms as capital letter sequences with more than two and fewer than five letters (with or without periods) that have at least five occurrences in the respective time span.[12] Forms with non-capital letter characters are excluded apart from plural forms with a small *s*. The design of suitable CQP queries takes into account that there is a multitude of possible underlying full forms and a high number of different reduction types. In order to obtain high retrieval effectiveness, the query results also contain a high number of irrelevant cases that have to be sorted out, e.g., letters denoting geometric objects, chemical symbols, and abbreviations of individual words (e.g., a rectangle *ABCD*, the chemical structure *COOH*, or the abbreviation *MSS*[13]) and regular words spelt with capital letters. The development of the scientific initialisms from the three time slices will be discussed on the basis of a quantitative and qualitative analysis. For each occurrence of the initialisms, the primary and secondary topics of the respective texts are extracted from *RSC* V6.0.4, as both topic types give equally valuable information on the

[11] In order to find out whether initialisms are used as parts of longer scientific terms in the *RSC* texts, the preceding and following tokens of the initialisms are also checked.

[12] Items that occur less frequently represent primarily other types of capital letter sequences or OCR errors. It is also more difficult to identify potential full forms of low-frequency forms as they are not distinctively linked to one specific term yet or occur only in one individual text.

[13] I.e. manuscripts.

content and scientific domains of the respective texts. For instance, a text with a usage of an initialism may have meteorology and geography or electricity and chemistry annotated as topics. The list of all these extracted topics will be visualised for each 30-year time interval with the word cloud function from the MATLAB Text Analytics Toolbox[14] to illustrate the development of the topics of the texts in which scientific initialisms are found. Finally, surprisal values for each occurrence of the initialisms are extracted from *RSC* V6.0.1. Their ranges will be compared for the three time spans. Surprisal, i.e. (un)predictability in context, serves as an indicator for cognitive processing effort (cf. Section 2.1; Shannon 1948; Hale 2001; Teich, Martínez Martínez & Karakanta 2020). Surprisal has been claimed to be proportional to the cognitive effort required to process any linguistic unit in different contexts of interaction and has been used in previous corpus studies to model and explain linguistic behaviour and choices (cf. Degaetano-Ortlieb & Teich 2022 for examples and a wider overview). Highly predictable linguistic units with low surprisal will require lower cognitive processing effort than less predictable linguistic units with higher surprisal (affecting, for instance, reading times). I will therefore analyse the surprisal values of scientific initialisms in the context of their preceding tokens in the *RSC*.

2.2 Analysis and Results

2.2.1 Frequencies

Most forms found between 1830 and 1919 were excluded from the query results as they either turned out to be no initialisms or they fell under other types of initialisms as described above. Among the 30 most frequent initialisms between 1830 and 1919, for instance, 21 are irrelevant for our purposes. The majority of the excluded initialisms in the data until 1919 denote people, their memberships, titles and ranks, providing information on the authors and communicators of the articles. This shows that the status and identities of the discourse participants played an important role in academic publications in the analysed time span. This information was encoded in relatively long nominal expressions so that conventionally used shortened forms have developed early, and many of them were already in usage before 1830. In the analysed time span, the authors and communicators

[14] https://www.mathworks.com/products/text-analytics.html (accessed 26 February 2024).

of the articles had access to similar types of social capital resources, being, for instance, *F.R.S.* ('Fellow of the Royal Society') and / or *F.L.S.* ('Fellow of the Linnaean Society'), *F.R.A.S.* ('Fellow of the Royal Astronomical Society') etc. The form *F.R.S.* occurs almost in every text. A relatively strong link between nobility and science and a tradition of orders and decorations is reflected in the use of initialisms after various author names, e.g., *K.C.S.I.* ('Knight Commander of the Star of India') or *K.C.M.G.* ('Knight Commander of St Michael and St George'). As expected, such initialisms for honorifics are most often found in headlines or by-lines of the articles, e.g., "*An Account of Experiments made with an Invariable Pendulum at New South Wales*, *by Major-General Sir Thomas Brisbane*, *K.C.B. F.R.S. Communicated by Captain Henry Kater*, *F.R.S.*, *in a Letter to Sir Humphry Davy*, *Bart. P.R.S.*"[15]. The majority of shortenings for multi-word terms in early scientific journal articles are thus mainly related to social culture and organisation. Some initialisms in the analysed time span refer to geographical multi-word expressions. Although such initialisms are also excluded from a more detailed analysis of *scientific* initialisms here, it is interesting to note that the *RSC* documents very early usages of geographical initialisms. For instance, *N.S.W.* ('New South Wales') is used in *RSC* texts from 1851 onwards, which antedates the *OED* quotations for this initialism starting in 1888 (*OED* s.v. *NSW*).

In the following, I will have a closer look at the initialisms for expressions related to scholarly topics discussed in the *RSC* journal articles. Only 10 types of initialisms for such expressions can be found between 1830 and 1859 (Tab. 2). The data from this time span yield some interesting insights into the early usages of scientific initialisms in LModE.

Tab. 2: Initialisms from specialised vocabulary in *RSC* texts from 1830–1859

	Initialism	**Freq. per 100,000 tokens**	**Full form**[16]
1	E.N.E.	0.56	east-north-east
2	W.N.W.	0.54	west-north-west
3	S.S.E.	0.45	south-south-east
4	N.N.W.	0.44	north-north-west
5	E.S.E.	0.43	east-south-east
6	N.N.E.	0.42	north-north-east

[15] *RSC* text ID: 107653, text year: 1833.

[16] In the *RSC* texts and in contemporary dictionary entries, e.g., in the *OED*, multi-word expressions that can be shortened by the use of initialisms are typically capitalised only for proper names (e.g., New South Wales). Otherwise, such full forms are mostly written in lowercase, although there are occasional variations. The short forms, however, almost always appear in uppercase letters.

	Initialism	Freq. per 100,000 tokens	Full form[16]
7	W.S.W.	0.24	west-south-west
8	S.S.W.	0.17	south-south-west
9	Q.E.D.	0.09	quod erat demonstrandum
10	N.P.D.	0.06	north polar distance
		Σ 3.04	

The 8 most frequent ones in this time span belong semantically together and refer to geographical information in nautical or meteorological terminology with a long tradition in specialised English. The underlying multi-word terms are attested already in English texts from the 14th century onwards (*OED* s.vv. *north-north-west*, *north-north-east*). Therefore it is not surprising that initialisms for them start to be used early in contexts where scholars would otherwise need to use the full terms repeatedly in their writing. These forms can also be found in LModE *RSC* texts before 1830.[17] *Q.E.D.* was also not newly coined in the time span we are looking at. It is the only initialism identified in the entire dataset that shortens a clausal structure. This short form has been used in English at least since the 15th century. The Latin full form 'quod erat demonstrandum' goes back to a Greek expression already used by mathematicians such as Euclid (c. 300 BC).

N.P.D. seems to be one of the first English initialisms shortening a nominal group compound (cf. Halliday & Martin 1993: 161). The full expression 'north polar distance' is attested in the *RSC* from the second half of the 18th century onwards. It reflects the new trend in LModE of forming scientific term patterns as nouns premodified by several lexical items such as other nouns or adjectives. The increasing usage of such clusters of lexical items in scientific lexemes in LModE generally led to more phrasal complexity and longer noun phrases in the *RSC* texts. On the one hand, this is counterbalanced by the introduction of initialisms for scientific multi-word terms that reduce the length of noun phrases (e.g., 'the degree in N.P.D.', 'changing the N.P.D. as required'). On the other hand, it leads to a compression of lexical information in noun phrases that makes it possible to pack even more information into such phrases by additional nominal pre- and postmodifiers. At the

[17] For example: "*The day was very fair and hot, with a little wind in the morning at W.S.W. which in the afternoon came round to N.N.W.*", *RSC* text ID: 105226, text year: 1757. Of course it is difficult to know whether such forms were pronounced as letter-by-letter initialisms. They may also have been used mainly as written abbreviations that would be read in their full forms as these were phonologically short with three syllables each.

end of the 19th century, long phrasal structures such as 'Greenwich N.P.D. observations of Polaris' or 'determination of the solar parallax from N.P.D. observations of Mars at Greenwich and Williamstown' have become commonly accepted in scientific articles. We can therefore observe from the 1860s onwards that scientific initialisms increasingly become shortcuts with wordhood status taking over the function of specialised vocabulary items consisting of multi-word units. These scientific multi-word units are typically premodified nominal groups with three lexical components. In the period from 1860–1889, the list of initialisms is semantically more diverse and longer than in 1830–1859, with 27 types and a higher proportion of nominal group compounds as full forms, e.g., *K.C.C.* 'kathodic closure contraction' (Tab. 3).

Tab. 3: Initialisms from specialised vocabulary in *RSC* texts from 1860–1889

	Initialism	**Freq. per 100,000 tokens**	**Full form**
1	E.M.F.(s)	1.69	electromotive force(s)
2	C.G.S.	0.68	centimetre-gramme-second
3	N.P.D.	0.31	north polar distance
4	W.S.W.	0.26	west-south-west
5	W.N.W.	0.23	west-north-west
6	E.N.E.	0.22	east-north-east
7	E.S.E.	0.20	east-south-east
8	N.N.W.	0.20	north-north-west
9	S.S.E.	0.19	south-south-east
10	N.N.E.	0.17	north-north-east
11	S.S.W.	0.16	south-south-west
12	G.M.T.	0.12	Greenwich mean time
13	A.C.C.	0.09	anodal[18] closure contraction
14	E.M.I.	0.09	electromotive intensity
15	K.C.C.	0.08	kathodic closure contraction
16	A.O.C.	0.07	anodal opening contraction
17	R.L.G.	0.06	rifle large grain
18	B.W.G.	0.05	Birmingham wire gauge
19	E.M.P.	0.05	electromagnetic pulse
20	K.O.C.	0.05	kathodic opening contraction
21	M.S.L.	0.05	mean sea level
22	N.G.F.	0.05	numerical generating function
23	C.E.M.F.	0.05	counter-electromotive force
24	E.M.E.	0.04	electromagnetic energy

[18] Or: anodic.

	Initialism	Freq. per 100,000 tokens	Full form
25	Q.E.D.	0.04	quod erat demonstrandum
26	L.M.T.	0.03	local mean time
27	R.G.F.	0.03	real generating function
		Σ 5.26	

The frequency of all scientific initialisms has risen from 3.04 to 5.26 per 100,000 tokens. Many are related to measurements and abstract terms. The frequency of the individual initialisms is not considerably higher than in the 30-year period before, apart from *E.M.F.*, the most frequent one with 1.69 occurrences per 100,000 tokens. Its structure and frequency have probably influenced the formation of similar shortenings in related domains in the same time span (*E.M.I.*, *E.M.P.*, *E.M.E.*, and the 4-letter form *C.E.M.F.*). *E.M.F.* behaves most lexeme-like among the observed forms and starts to take a plural suffix from the 1880s onwards. It is the only scientific initialism identified with an affix between 1830 and 1919 in the *RSC*. Generally, the one-token format gives initialisms a higher syntactic flexibility than their underlying MWU so that these short forms become regularly used as noun premodifiers from the 1870s onwards (e.g., *C.G.S. system*, *R.L.G. powder*, *B.W.G. diameter*).[19] Among those with lower frequencies, some were coined by the authors and introduced in the *RSC* texts as in the following example from an article by the mathematician Arthur Cayley from the 1870s:

> But the whole plan of the Memoir was changed by Sylvester‘s discovery of what I term the Numerical Generating Function (N.G.F.) of the covariants of the quintic, and my own subsequent establishment of the Real Generating Function (R.G.F.) of the same covariants. (*RSC* text ID: rspl_1878_0080)

For the majority of scientific initialisms in 1860–1889, the full expression is not used in the *RSC* in this time span. This indicates that the initialisms were already conventionally used in other forms and media of scientific discourse (e.g., spoken academic discourse or books), and they were in many cases not coined in the *RSC* journal texts. Sometimes, their underlying full expression is not used at all in the *RSC* between 1830 and 1919.

From 1890 to 1919, the number of scientific initialisms with at least five occurrences, in absolute numbers, has risen to 38 types (Tab. 4).

[19] Usages in other word-formation processes, e.g., as parts of adjective compounds such as *PVC-lined* or in derivations with prefixes such as *anti-BSA response*, can be found only much later in the *RSC*: from the 1950s onwards.

Tab. 4: Initialisms from specialised vocabulary in *RSC* texts from 1890–1919

	Initialism	**Freq. per 100,000 tokens**	**Full form**
1	E.M.F.	7.28	electromotive force
2	G.M.T.	5.81	Greenwich mean time
3	C.G.S.	2.85	centimetre-gramme-second
4	S.S.N./SSN	0.83	standard scale number
5	P.W.B.C.	0.43	polynuclear white blood corpuscles
6	M.L.D.	0.22	minimum lethal dose
7	S.B.P.	0.19	sulphur boiling point
8	R.B.C.	0.17	red blood corpuscles
9	N.T.P.	0.16	normal temperature and pressure
10	M.H.D.	0.15	minimum haemolytic dose
11	W.F.P.	0.12	water freezing point
12	A.C.E.	0.11	alcohol, chloroform, ether
13	S.W.G.	0.11	standard wire gauge
14	R.M.S.	0.11	root-mean-square
15	E.S.E.	0.11	east-south-east
16	S.S.E.	0.11	south-south-east
17	N.N.E.	0.10	north-north-east
18	W.B.P.	0.10	water boiling-point
19	E.N.E.	0.09	east-north-east
20	B.W.G.	0.09	Birmingham wire gauge
21	S.S.W.	0.08	south-south-west
22	L.C.M.	0.08	least common multiple
23	E.S.U.	0.07	electrostatic unit(s)
24	W.S.W.	0.07	west-southwest
25	E.M.E.	0.07	electromagnetic energy
26	N.N.W.	0.07	north-north-west
27	W.N.W.	0.06	west-north-west
28	B.O.T.[20]	0.05	Board of Trade
29	E.M.P.	0.04	electromagnetic pulse
30	M.S.L.	0.04	mean sea level
31	B.T.U.	0.03	Board of Trade unit
32	A.V.B.	0.03	atrio-ventricular bundle
33	M.D.W.	0.03	Mather-Duddell wattmeter
34	N.P.D.	0.03	north polar distance
35	Q.E.D.	0.02	quod erat demonstrandum
36	L.M.T.	0.02	local mean time
37	A.C.C.	0.02	anodal closure contraction
38	R.L.G.	0.02	rifle large grain
		Σ 19.97	

[20] As indicated above, this initialism for a government body was not excluded when it was used in the term ‘B.O.T. cell(s)’.

The overall frequency of these forms has risen considerably to 19.97 per 100,000 tokens. New forms have been coined from a greater variety of topics. The proportion of texts with at least one scientific initialism increases considerable over time. Generally the article content has also become much longer over time with more opportunities and requirements for reductions. Almost all identified initialisms still contain periods. The first variant without periods is used from 1900 onwards (*SSN* for 'standard scale number'). Short forms for lists of nouns as premodifiers also become common (e.g., *A.C.E. mixture*). Among the top five forms, we have a 4-letter initialism, *P.W.B.C.*

Acronymic initialisms, particularly those that resemble existing words, seem to be of marginal importance in the analysed time span. They do not seem to be an innovation of scientific language as the few that can potentially be pronounced like words typically come from other semantic categories. *B.O.T. cells*, for instance, contains an initialism that stands for a government body. Among the ones that were excluded from the results for semantic reasons, we also find examples such as *M.I.C.E.* ('Member of the Institution of Civil Engineers') and *M.A.P.S.* ('Member of the American Philosophical Society') in texts from the second half of the 19th century.

2.2.2 Topics

The word clouds in Fig. 1–3 show an overview of the development of the topics of the texts in which the scientific initialisms were used. The topics were extracted from *RSC* V6.0.4 for each usage of the scientific initialisms listed in Tab. 2–4 and visualised with the MATLAB Text Analytics Toolbox.

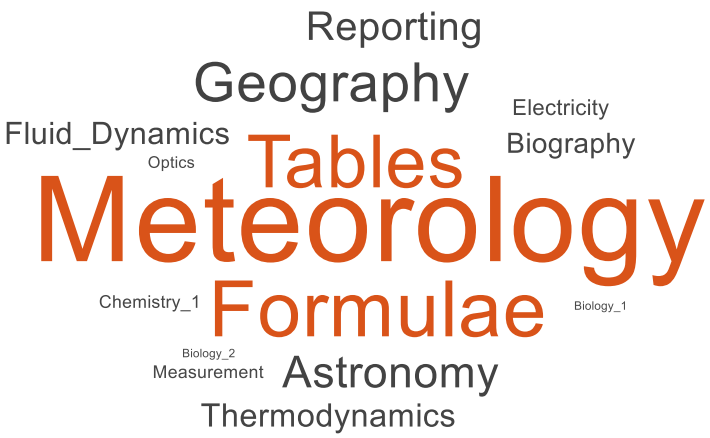

Fig. 1: Topics of texts in which initialisms were used (1830–1859)

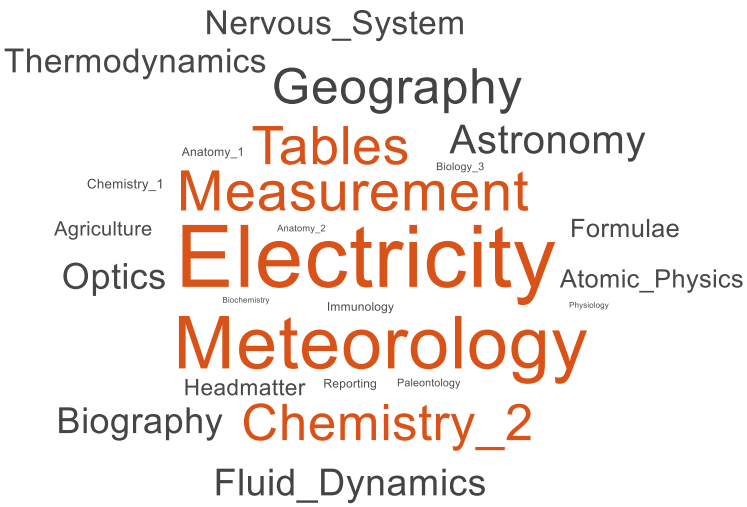

Fig. 2: Topics of texts in which initialisms were used (1860–1889)

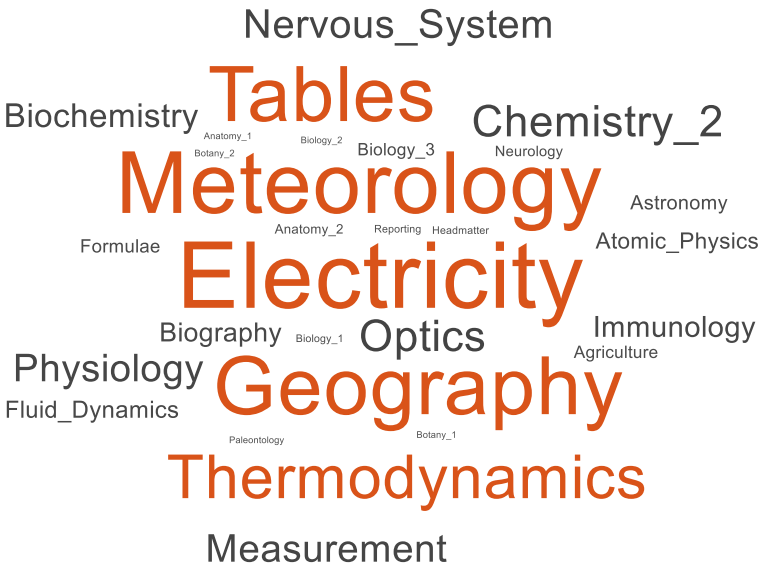

Fig. 3: Topics of texts in which initialisms were used (1890–1919)

The dominant topics from the first time span were meteorology, geography, astronomy, and mathematical contexts (formulae, tables). Similar topics remain among the most important ones in the following times spans, but initialism start to represent a greater variety of the topics that are covered in an increasingly specialised way in the scientific journal texts in the *RSC*, e.g., the sciences of electricity or biochemistry. Texts from all topics covered in the *RSC* contain initialisms in the third time span. However, the language used with regard to some topics that are represented by a non-negligible number of texts in the *RSC* is not characterised by a high number of initialisms, particularly in the biological sciences (e.g., anatomy, botany, and physiology).

2.2.3 Surprisal

Fig. 4 presents an overview of the surprisal ranges of the individual usages of the scientific initialisms in the three analysed time spans to measure the (un)predictability of the initialisms given their preceding textual contexts.

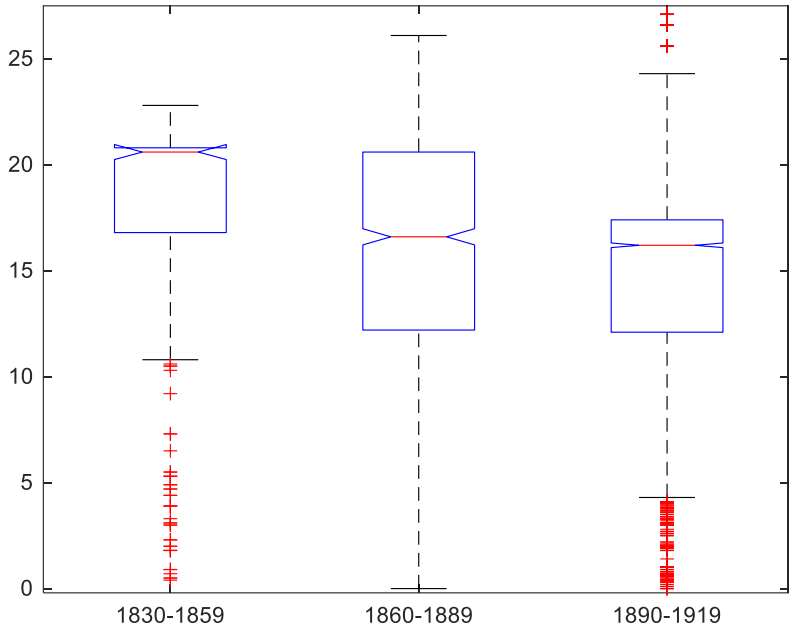

Fig. 4: Surprisal values of initialisms in the *RSC* from 1830–1919

The observed surprisal values of scientific initialisms in their textual contexts decrease over time. The median in the first time span (20.6) is significantly higher than in in the second and third time span (16.6 and 16.2), which we can conclude from the plotted notches that represent the confidence interval around the medians. The 1830–1859 period has generally high surprisal values in a relatively small interquartile range IQR. The third period from 1890–1919 has generally lower surprisal values in a similarly small IQR. The 75th percentile values in the first two time spans are very similar to each other (20.8 and 20.6), while the 25th percentile values from the second and third time span are also very similar to each other (12.2 and 12.1). The second time span therefore has a higher IQR and seems to represent a transitional period between the first and the third one with regard to the surprisal development.

The decreasing surprisal of scientific initialisms can be explained by their increasing occurrence in fixed contexts. In the second and third time span, initialisms are more regularly preceded and followed by certain words, e.g., when they are used as nominal

premodifiers as in the examples discussed above or when they are preceded by certain modifiers or sequences of words, e.g., *Atlantic M.S.L.*, *temporary E.M.I.*, *the rise of E.M.F.*, *magnetic field in C.G.S.*

3. Summary and Outlook

The results of the analysis of the *RSC* texts show that initialisms for scientific concepts from the mathematical, physical, and life sciences became common shortening devices in scientific articles during the analysed time span. The overall frequency of these forms has risen considerably between 1830 and 1919, particularly in the context of the ongoing specialisation of the sciences during the final period of LModE and at the beginning of PDE. The scientific initialisms that were identified shorten and replace multi-word nominal expressions that had become conventionalised scientific terms. Initialisms are visually distinct indicators of terminology that are used across texts from the same specialised domains and in individual texts in repetition-based lexical chains. This has further consequences for the word-formation system of English with implications for other linguistic levels. Initialisms have established themselves as an innovative strategy in specialised contexts and as one of the linguistic means of the English language that lead to maximally compact and informationally dense units and efficient expert-to-expert communication. The regular usage of initialisms for scientific multi-word terms systematically reduces the length of noun phrases. It also leads to a compression of lexical information in noun phrases that makes it possible to pack even more information into phrasal structures with scientific content. We observe particularly from the 1860s onwards that scientific initialisms have increasingly become shortcuts for multi-word units with wordhood and term status.

From the analysis of the full forms of the initialisms and the text topics extracted from the textual metadata, I conclude that early scientific initialisms in the *RSC* are mainly related to mathematical contexts and measurements and to nautical, meteorological, and astronomical terminology. In various cases, these short forms are still used in PDE. During the analysed period, initialisms represent an increasing diversity of specialised fields, e.g., the sciences of electricity or biochemistry. Some fields within the biological sciences were

still characterised by a rather low frequency of initialisms. The surprisal values of scientific initialisms decreased over time as these initialisms increasingly occurred in more conventionalised textual contexts and fixed phrasal expressions. Overall, the corpus analysis of the scientific periodicals of the Royal Society of London shows that key developments towards the conventionalisation of scientific initialisms as term formation patterns took place between 1830 and 1919. Comparisons with other data from the same time span will probably confirm that the results reflect a general development in scientific English, but due to the inherent limitations of diachronic specialised resources, it could be argued that there are certain characteristics that may bias the results.

One important finding of the analysis of the *RSC* is that the one-token format has given initialisms a higher syntactic flexibility than their underlying MWU. They have become regularly used as noun premodifiers, they started to take inflectional suffixes and increasingly occurred in variants without periods in the analysed time span. An analysis of the *RSC* texts after 1920 will reveal an even greater diversity of initialisms as shortening strategies, an even more lexeme-like behaviour of these forms and an increasing number of initialisms undergoing further word-formation processes. From the 19th century onwards, initialisms have acquired more functions and features than mere abbreviations, and they therefore deserve a much more prominent role in contemporary morphological theory.

References

Algeo, John. 1975. The Acronym and its Congeners. In Adam & Valerie Makkai (eds.), *Proceedings of the 1st LACUS Forum 1974*, 217–234. Columbia, SC: Hornbeam Press.

ARCHER-3.2 = Representative Corpus of Historical English Registers version 3.2. 2013. Originally compiled under the supervision of Douglas Biber & Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. https://www.projects.alc.manchester.ac.uk/archer/ (accessed 26 February 2024).

Barnett, Adrian & Zoe Doubleday. 2020. Meta-Research: The Growth of Acronyms in the Scientific Literature. *eLife 9*: e60080. DOI: 10.7554/eLife.60080.

Bauer, Laurie. 1983. *English Word-Formation*. Cambridge: Cambridge University Press.

Baum, S. V. 1962. The Acronym, Pure and Impure. *American Speech* 37(1). 48–50.

Berkwits, Michael. 2003. Capture! Shock! Excite! Clinical Trial Acronyms and the 'Branding' of Clinical Research. *Annals of Internal Medicine* 133(9). 755–762.

Biber, Douglas & Bethany Gray. 2016. *Grammatical Complexity in Academic English*: *Linguistic Change in Writing*. Cambridge: Cambridge University Press.

Grange, Bob & D. A. Bloom. 2000. Acronyms, Abbreviations and Initialisms. *BJU International* 86(1). 1–6. DOI: 10.1046/j.1464-410x.2000.00717.x.

Cannon, Garland. 1989. Abbreviations and Acronyms in English Word-Formation. *American Speech* 64(2). 99–127.

Crespo, Begoña & Isabel Moskowich. 2015. A Corpus of History Texts (CHET) as Part of the Coruña Corpus Project. *Proceedings of the International Scientific Conference* "*Corpus linguistics – 2015*". 14–23.

Degaetano-Ortlieb, Stefania & Elke Teich. 2022. Toward an Optimal Code for Communication: The Case of Scientific English. *Corpus Linguistics and Linguistic Theory* 18(1). 175–207.

Dietz, Klaus. 2015. Historical Word-Formation in English. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Volume 3*: *Word-Formation*: *An International Handbook of the Languages of Europe*, 1914–1930. Berlin & New York: De Gruyter Mouton.

Evert, Stefan. 2005. *The CQP Query Language Tutorial.* IMS: Stuttgart University.

Fandrych, Ingrid. 2004. Non-Morphematic Word-Formation Processes: A Multi-Level Approach to Acronyms, Blends, Clippings and Onomatopoeia. PhD dissertation. Bloemfontein: University of the Free State.

Fandrych, Ingrid. 2008. Submorphemic Elements in the Formation of Acronyms, Blends and Clippings. *Lexis* (2). 105–123.

Fankhauser, Peter, Jörg Knappen & Elke Teich. 2016. Topical Diversification over Time in the Royal Society Corpus. In Maciej Eder & Jan Rybick (eds.), *Digital Humanities 2016 Conference Abstracts*, 496–500. Kraków: Jagiellonian University & Pedagogical University.

Fischer, Stefan, Jörg Knappen, Katrin Menzel & Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. In *Proceedings of the 12th Language Resources and Evaluation Conference* (*LREC*), 794–802. Marseille: European Language Resources Association.

Fred, Herbert L. & Tsung O. Cheng. 2003. Acronymesis: The Exploding Misuse of Acronyms. *Texas Heart Institute Journal* 30(4). 255–257.

González, Félix Rodríguez & Garland Cannon. 1994. Remarks on the Origin and Evolution of Abbreviations and Acronyms. In Francisco Fernández, Miguel Fuster Márquez & Juan Jose Calvo (eds.), *English Historical Linguistics 1992*: *Papers from the 7th International Conference on English Historical Linguistics*, *Valencia*, *22–26 September 1992*, 261–272. Amsterdam: Benjamins.

Hale, John. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (*NAACL'01*). 1–8. Pittsburgh, PA: Association for Computational Linguistics.

Halliday, Michael A. K. & James R. Martin. 1993. *Writing Science Literacy and Discursive Power*. 2nd edition. London: Falmer Press.

Hardie, Andrew. 2012. CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics* 17(3). 380–409. DOI: 10.1075/ijcl.17.3.04har.

Heller, Louis G. & James Macris. 1968. A Typology of Shortening Devices. *American Speech* 43(3). 201–208.

Hitchcock, Tim, Robert Shoemaker, Clive Emsley, Sharon Howard & Jamie McLaughlin et al. (eds.), *The Old Bailey Proceedings Online*, 1674–1913. https://www.oldbaileyonline.org, version 7.0 (accessed 26 February 2024).

Hock, Hans Henrich. 2021. *Principles of Historical Linguistics*. 3rd revised and updated edition. Berlin & Boston: De Gruyter Mouton.

Kermes, Hannah, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen & Elke Teich. 2016. The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation* (*LREC*). 1928–1931. Portorož: European Language Resources Association (ELRA).

Kreidler, Charles W. 1979. Creating New Words by Shortening. *English Linguistics* 13(1). 24–36.

Marchand, Hans. [1960] 1969. *The Categories and Types of Present-day English Word-Formation.* 2nd edition. Munich: C.H. Beck.

Mattiello Elisa. 2012. Abbreviations in English and Italian Scientific Discourse. *ESP Across Cultures* 9. 149–168.

Mattiello, Elisa. 2013. *Extra-Grammatical Morphology in English*: *Abbreviations*, *Blends*, *Reduplicatives and Related Phenomena*. Berlin & Boston: De Gruyter Mouton.

Menzel, Katrin, Jörg Knappen & Elke Teich. 2021. Generating Linguistically Relevant Metadata for the Royal Society Corpus. *Research in Corpus Linguistics* 9(1). 1–18.

OED = The Oxford English Dictionary. 2000–. 3rd edition. http://www.oed.com (accessed 26 February 2024).

Rogers, Walter T. 1913. *Dictionary of Abbreviations.* London: George Allen & Company, LTD.

RSC = Royal Society Corpus. V6.0.1 and V6.0.4. https://fedora.clarin-d.uni-saarland.de/rsc_v6/; https://corpora.clarin-d.uni-saarland.de/cqpweb/ (accessed 26 February 2024).

Shannon, Claude E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27(3). 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Tay, Andy. 2020. Snappy Acronyms Generate Excitement for Science (SAGES). *The Scientist*. https://www.the-scientist.com/news-opinion/snappy-acronyms-generate-excitement-for-science--sages--67057 (accessed 26 February 2024).

Teich, Elke, José Martínez Martínez & Alina Karakanta. 2020. Translation, Information Theory and Cognition. In Alves, Fabio & Arnt Lykke Jakobsen (eds.), *The Routledge Handbook of Translation and Cognition*, 360–375. London: Routledge.

Katrin Menzel
Universität Mannheim
Anglistik IV
D–68159 Mannheim
katrin.menzel@uni-mannheim.de