http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Algorithmic Complexity and Learnability in German Monolingual Learner Lexicography. A Case Study

Alberto Galván-Santana, *Department of Applied Linguistics,
Universitat Politècnica de València (UPV), Spain
(algalsan@doctor.upv.es) (https://orcid.org/0000-0002-4665-5956)*

Abstract: This paper analyzes the algorithmic complexity (also known as Kolmogorov complexity or descriptive complexity) of the lemma corpus included in the *Wortfamilienwörterbuch der deutschen Gegenwartssprache* (WfWG; Augst 2009) as a function of its macrostructural arrangement. The results show that, compared to the alphabetical order, the WfWG word-family arrangement produces an algorithmically more compressible, and therefore less complex version of the lemma corpus. This observation points to a higher degree of learnability and cognitive accessibility of the lemma corpus arranged in word families.

Keywords: MONOLINGUAL LEARNER'S DICTIONARY, MACROSTRUCTURE, NAVIGATION, LEARNABILITY, ALGORITHMIC COMPLEXITY, COMPRESSION

Zusammenfassung: Algorithmische Komplexität und Lernbarkeit in der einsprachigen Lernerlexikographie des Deutschen. Eine Fallstudie. Dieser Beitrag analysiert die algorithmische Komplexität (auch Kolmogorow-Komplexität oder Beschreibungskomplexität) des im *Wortfamilienwörterbuch der deutschen Gegenwartssprache* (WfWG; Augst 2009) enthaltenen Lemmakorpus in Abhängigkeit von dessen makrostruktureller Anordnung. Die Ergebnisse zeigen, dass die Wortfamilienanordnung im Vergleich zur alphabetischen Reihenfolge eine stärker komprimierbare und daher weniger komplexe, d. h. kürzer beschreibare Version des WfWG-Lemmakorpus darstellt. Diese Beobachtung deutet auf einen höheren Grad an Lernbarkeit und kognitiver Zugänglichkeit des in Wortfamilien angeordneten Lemmakorpus hin.

Schlüsselwörter: EINSPRACHIGES LERNERWÖRTERBUCH, MAKROSTRUKTUR, NAVIGATION, LERNBARKEIT, ALGORITHMISCHE KOMPLEXITÄT, KOMPRESSION

1. Introduction

The increasing use of the monolingual learner's dictionary (MLD) in the field of L2 acquisition has created a growing demand for a psycho-cognitive motivation in the conception of lexicographic texts that would facilitate an intrinsic activation of grammatical knowledge in the dictionary user (Fuertes-Olivera and

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Tarp 2011; Haß-Zumkehr 2012; Kövecses and Csábi 2014; Kremer et al. 2008).

Nowadays the psycho-cognitive approach in the elaboration of MLD is based on different models and cognitive theories focused on semantic aspects: frame semantics (Fillmore 1982, 1985), the theory of conceptual metaphor (Lakoff 1993; Lakoff and Johnson 1980), or the model of principled polysemy (Evans 2009; Tyler and Evans 2004), among others. With regard to the microstructure, recent approaches have led to the formulation of a "cognitive lexicography" (Ostermann 2015), focused on the applicability of cognitive linguistics to the propositional format of the lemma definition. Regarding the macrostructure, the application in the area of e-lexicography of these linguistic-cognitive theoretical frameworks has resulted in the creation of lexical knowledge databases such as WordNet (Miller et al. 1990), MindNet (Richardson et al. 1998), FrameNet (Fillmore et al. 2003), and HowNet (Dong and Dong 2006).[1] In print lexicography, the application of various linguistic knowledge and theories to lexicography has produced a variety of macrostructural designs whose intended goals can be reduced to a single common denominator: assisting the L2 student in the acquisition of language production and comprehension skills. This (primarily) printed dictionary type is represented by onomasiological dictionaries (Casares 2013; Dornseiff 2020; Rey-Debove and Rey 2009; Simone 2010), collocation and combinatorial dictionaries (Benson et al. 2009; Bosque 2006; Häcki Buhofer et al. 2014; Mel'čuk et al. 1999), and word-family dictionaries (Augst 2009; Davau et al. 1984; Kirkpatrick 1983), among others.[2]

In this context, the general objective of this project is to promote the development of macrostructural arrangement criteria capable of facilitating language acquisition in MLD users through their involvement in cognitive-efficient inference processes. This project is thus conceived as a contribution aiming to close the aforestated research gap — the demand for psycho-cognitive approaches in monolingual learner lexicography — with the lexicographic macrostructure at the center of the analysis.

To this end, we propose to address the psycho-cognitive relevance of the macrostructure from an extralinguistic perspective: the Algorithmic Information Theory (AIT; Chaitin 2004; Grünwald and Vitányi 2008). This information-based approach will allow us to analyze the cognitive accessibility of the lemma corpus in terms of algorithmic complexity (AC; Kolmogorov 1963; Chaitin 1969; Solomonoff 1964a, 1964b). The notion of AC is inversely correlated with learnability so that the learnability of any given data set increases with decreasing complexity of its structural organization (Clark and Lappin 2013; Fulop and Chater 2013; Kempe et al. 2015; Zenil and Gauvrit 2017).[3] In this regard, we argue that language learning with the help of a (monolingual learner's) dictionary, in this case the *Wortfamilienwörterbuch der deutschen Gegenwartssprache* (WfWG), can generally be conceived of in computational terms as a supervised learning task, in which lower complexity of the data structure (the lemma corpus) expedites the path for the learning algorithm (the dictionary user) to efficiently approximate the program (the grammar) that generates the data (cf. Grünwald 2005: 7-10).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

In accordance with this methodological approach, the overall objective of this study translates into the following specific objectives:

(O1) determine whether the macrostructural arrangement of the WfWG lemma corpus has an impact on its AC value;

(O2) evaluate to what extent the AC value of the WfWG corpus varies as a function of its macrostructural order, in word families and alphabetical, respectively;

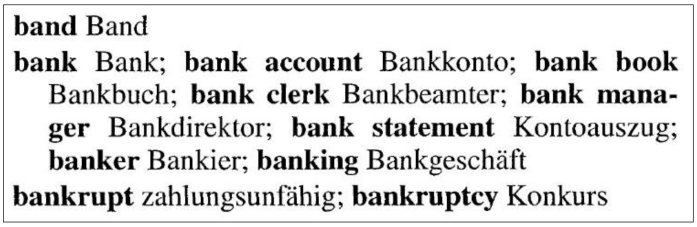
(O3) identify, among the aforementioned ordering criteria, the macrostructural arrangement leading to a lower AC value to the corpus.

1.1 The lexicographic macrostructure

In general terms, the word macrostructure refers to the external structure (of a vertical or paradigmatic nature) that relates to the lemma corpus and the ordered representation of its elements (Engelberg and Lemnitzer 2009). The most elementary version of an alphabetic macrostructure is the simple alphabetical order (Gouws 2003). In a simple alphabetic dictionary, the number of indexed items, which are always main lemmata, is equal to the number of dictionary articles (Wiegand 1989). This implies that all lemmata are ordered according to the alphabetic value of the first character of the lemma sign and, consecutively, according to the alphabetic value of the following characters. Martínez de Sousa (2009: 214) defines this lexicographic procedure as "simple or lexicological" alphabetical arrangement.

In this regard, an alternate alphabetical arrangement procedure is characterized by the presence of groupings of sublemmata (Gouws 2003). These groups — integrated and hierarchically subordinated to the main lemmata — result from the inclusion of complementing lemmata that, in contrast to the main paradigmatic arrangement, show a syntagmatic composition as a single textual block. This block, accessed through the main lemma, is composed of the articles associated with each of the subsequent sublemmata (Gouws 2003). Such clusters of sublemmata can be classified into two different categories, niches and nests, depending on whether they adhere to the prevailing alphabetical order (Bergenholtz and Tarp 1995; Gouws 2003; Hausmann and Wiegand 1989; Wiegand 1989).

As Figure 1 shows, the sublemmata grouped in niches adhere to an alphabetical order, not only in their internal (horizontal) organization but also concerning the external (vertical) arrangement of the main lemma corpus. In this way, the sublemmata integrated within the niche, in addition to showing an internal alphabetical order, alphabetically precede the following main lemma: "This type of cluster merely illustrates a deviation in the direction of macrostructural ordering, i.e. horizontal instead of vertical, but does not imply any deviation from the prevailing straight alphabetical ordering" (Gouws 2003: 41).



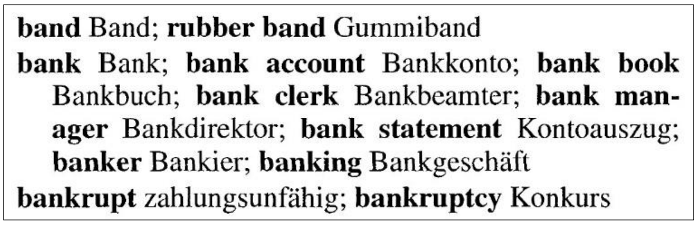
band Band

bank Bank; **bank account** Bankkonto; **bank book** Bankbuch; **bank clerk** Bankbeamter; **bank manager** Bankdirektor; **bank statement** Kontoauszug; **banker** Bankier; **banking** Bankgeschäft

bankrupt zahlungsunfähig; **bankruptcy** Konkurs

Figure 1: Illustrative example of a niche grouping procedure in an English–German bilingual dictionary (Bergenholtz and Tarp 1995)

Nesting differs from niche grouping in one important aspect: although the sublemmata included in the nest follow the preceding main lemma alphabetically, the nest includes sublemmata that do not conform with the alphabetical value of the succeeding main lemma: "[A]s opposed to niching, nesting enables interruption within the nest of the order of graphemes in the access alphabet in order that all lemmata with the same stem may appear together in the same article" (Bergenholtz and Tarp 1995: 194). This deviation from the strict alphabetical order is shared by the two types of nesting: first-level and second-level nesting. As displayed in Figure 2, first-level nesting represents an intermediate stage between the niche and the second-level nesting: the arrangement of sublemmata in this first-level nesting obeys to a strict alphabetical order, however, some sublemmata alter this maxim concerning the following main lemma.



band Band; **rubber band** Gummiband

bank Bank; **bank account** Bankkonto; **bank book** Bankbuch; **bank clerk** Bankbeamter; **bank manager** Bankdirektor; **bank statement** Kontoauszug; **banker** Bankier; **banking** Bankgeschäft

bankrupt zahlungsunfähig; **bankruptcy** Konkurs

Figure 2: Illustrative example of first-level nesting in an English–German bilingual dictionary (Bergenholtz and Tarp 1995)

This strict internal alphabetical arrangement of the first level nesting constraints its lexicographic functionality as compared with the degree of "sophistication" of the second level nesting (Gouws 2003: 41), characterized by a further internal cancellation of the alphabetical order, as will be seen in the WfWG: "[S]econd level nesting gives evidence of a lexicographic procedure where morphosemantic motivations dominate the alphabetical ordering principle in the presentation of sublemmata in a horizontal lemma file" (ibid.: 43).

1.2 The macrostructure of WfWG

The *Wortfamilienwörterbuch der deutschen Gegenwartssprache* is the first and only one-volume learner's dictionary of contemporary German whose macrostructure is organized around word families (Augst 2009).[4] According to Augst (ibid.), this arrangement in word families plays a fundamental role in the acquisition of German (both L1 and L2) to the extent that it facilitates an easy comprehension of the internal mechanisms governing word formation and, by extension, of the more general patterns that connect and organize the lexical units at a higher order level: the grammar.[5]

Augst's proposal (1992: 34) in the WfWG arises from an approach, according to which "in der Wortbildung selbst (wie aber auch in der Wortbildungstheorie) Produktivität auf Grund genereller Regeln und Produktivität auf Grund singulärer Analogie nebeneinander (be)stehen [in the very practice of lexical formation (as well as in the theory of lexical formation) coexist both rule-based and analogy-based productivity; our translation]".[6] On this basis, Augst proposes that, in order to appropriately recreate the lexicon's word-family structure, its lexicographic representation must conform to the "relative motivation" (*relative Motiviertheit*; 2009: IX) between the formal manifestations of the lexical units at a given moment. The starting point of Augst's lexicographic approach lies in the "synchronous etymological competence" (*synchrone etymologische Kompetenz*; 1975: 156-231) understood as any speaker's perceived ability to motivate lexical relations. This ability entails decomposing and reducing the lexicon complexity for the purpose of filtering the morphological core that conveys the central lemma element of a word family (Augst 2009). This scheme, illustrated in Figure 3, facilitates the tracing of an itinerary in the opposite direction towards word formation processes, and the construction, in accordance with the successive derivations (of 1st, 2nd, 3rd degree, etc.), of a hierarchical and recursive or replicating lexical structure (Augst 2009).

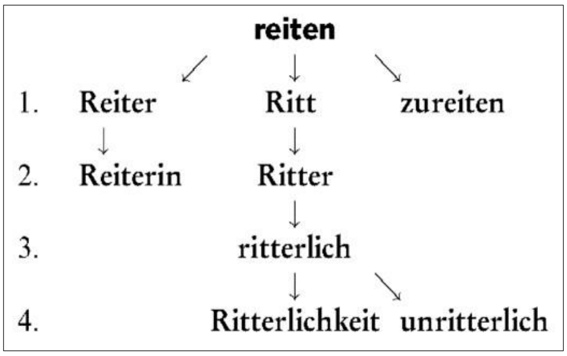
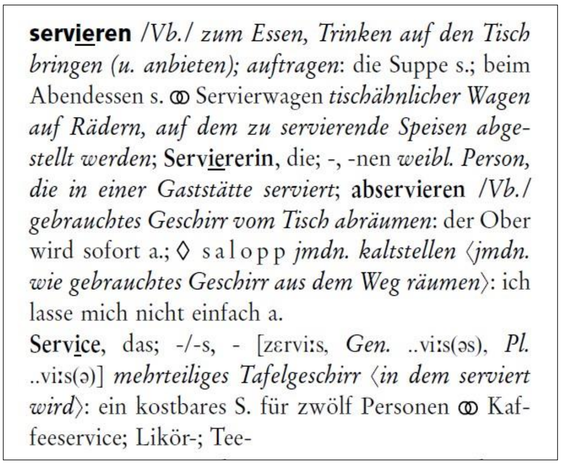


Figure 3: Example of a hierarchical and recursive structure representing the word family *reiten* (Augst 2009)

The arrangement of the lemma corpus according to this type of structure results in a second-level nesting macrostructure (Figure 4). Each headword (*Kernwort*) of the respective word family is listed as a main lemma. Following the headword, the first-degree derivations are listed as sublemmata in independent paragraphs; first, the suffixed derivations, followed by the prefixed derived forms. Second to nth-degree derivations are added horizontally to the first-degree derivations in the corresponding paragraph. The compounds are located after the pertinent sublemma and are identified through the indicator (ↀ). The compounds are arranged in such a manner that those compounds in which the sublemma appears as the modifier or *determinans* (*Bestimmungswort*) are listed first. Next, an en-dash precedes compounds in which the sublemma serves as the head or *determinatum* (*Grundwort*). If there is a linking element (*Fugenelement*) between the compound constituents, the sublemmata are sorted alphabetically according to this linking element (for example, compounds with *Wort* have either no linking element or *-er-*, which results in the following arrangement: *Wort* ... ↀ *Wortart*; *-gruppe*; *-schatz* — *Wörterbuch* — *Beiwort*). The compounds can, in turn, provide the lexical base for additional second-degree compounds, etc. These compounds appear in parentheses (for example: *Wörterbuch* … (*Bild-*; *Fach-*; *Hand-*).



servieren */Vb./ zum Essen, Trinken auf den Tisch bringen (u. anbieten); auftragen*: die Suppe s.; beim Abendessen s. ↀ Servierwagen *tischähnlicher Wagen auf Rädern, auf dem zu servierende Speisen abgestellt werden*; **Serviererin**, die; -, -nen *weibl. Person, die in einer Gaststätte serviert*; **abservieren** */Vb./ gebrauchtes Geschirr vom Tisch abräumen*: der Ober wird sofort a.; ◊ salopp *jmdn. kaltstellen* ⟨*jmdn. wie gebrauchtes Geschirr aus dem Weg räumen*⟩: ich lasse mich nicht einfach a.

Service, das; -/-s, - [zɛrviːs, *Gen.* ..viːs(əs), *Pl.* ..viːs(ə)] *mehrteiliges Tafelgeschirr* ⟨*in dem serviert wird*⟩: ein kostbares S. für zwölf Personen ↀ Kaffeeservice; Likör-; Tee-

Figure 4: Section of the entry related to the lemma *servieren* (Augst 2009) featuring a second-level nesting configuration in the macrostructure

The lexicographic (re)production of this organizing principle reproduces, according to Augst (2009), the structuring of the lexicon around ideal archetypes as perceived by the speaker by virtue of his linguistic competence. In this regard, the conceptual challenge of Augst's lexicographic endeavour lies in "linearizing" the grammatical intricacy of word families into the macrostructure (2009: XII), in such a way that the dictionary user will be able to

(i) recognize the word family hierarchical structure despite the "linearization problem" (Geeraerts 2001: 18) inherent to printed dictionaries, and

(ii) find without difficulty the word that prompted the search, "[d]abei soll die Wortfamilienstruktur für die Bedeutungsangaben jedes einzelnen Wortes der Wortfamilie wechselseitig erhellend wirken und somit die Zerrissenheit des alphabetisch-semasiologischen Wörterbuchs aufheben [yet the word family structure must have a reciprocal highlighting effect on the meaning of each word in the family and thus neutralize the disintegration [of morphosemantic relationships] of the alphabetic-semasiological dictionary; our translation]" (Augst 2009: IX).

2. Material and method

In the following analysis, the macrostructural ordering type is the independent variable whose modification produces an alteration in the corpus AC value as the dependent variable. The test sets will be composed of a relevant lemma collection extracted from the WfWG corpus and subjected to the respective macrostructural arrangements according to (i) the original WfWG arrangement in word families and (ii) the alphabetical arrangement of said corpus. The control set consists of a disordered macrostructure having a random distribution of the same corpus elements.

2.1 Material

The lemma corpus used to perform this study will be extracted from the WfWG (Augst 2009). The lemma selection process to build this corpus will focus on finding "a small and insightful subset" of the original corpus (David et al. 2016). To this end, a double criterion is established: (i) morphological productivity, and (ii) the presence of a lexicographic definition associated with the lemma. According to these parameters, the lemma corpus will consist of the whole set of headwords (including homonyms) and their derived forms. Since word compounding has limited grammatical relevance on a synchronic level, compounds will be excluded. The result generates a set $L = \{l_1, l_2, l_3, \dots l_{m-1}, l_m\}$ of 27,622 lemmata. This set is treated and presented as a string of n characters belonging to $C = \{c_1, c_2, c_3, \dots c_{n-1}, c_n\}$, where c_i represents any character $c \in C$ at position i. This procedure results in a string comprising 261,121 characters.

2.2 Method

Algorithmic complexity being formally incomputable, the application of the Minimum Description Length principle (Grünwald 2005, 2007; Rissanen 1978)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

will allow us to obtain an estimate of the corpus AC value: "[t]he goal of statistical inference may be cast as trying to find regularity in the data. 'Regularity' may be identified with 'ability to compress.' [Minimum Description Length] combines these two insights by *viewing learning as data compression* [emphasis in the original]" (Grünwald 2007: 12). In other words, the more compressed the data set, the greater the extent to which a system can be said to have learned on that set (Chaitin 2006; Maguire et al. 2015).

In accordance with this methodological framework, the following experimental design is based on the study conducted by Koplenig et al. (2017) on the statistical correspondence between the internal structuring of the lexicon and its syntagmatic ordering. The adopted notations as well as the ensuing explanations derive from said study.

Our interest is focused on determining the amount of regularity of the lemma corpus as a variation of its entropy rate (Koplenig et al. 2017). On this basis, the absolute redundancy (D) will serve as the reference magnitude. This magnitude measures the difference between the absolute rate of entropy (R_0) — that is, its maximum value — and the real or effective rate of entropy (r), being a high value of D indicative of a greater amount of regularity in the set (Koplenig et al. 2017).

$$D = R_0 - r \tag{1}$$

In order to isolate the amount of absolute redundancy (D) that can be attributed to the different lemma arrangements, we will first estimate the entropy value of the control set, that is, of the set L in randomized order. This value corresponds to the absolute entropy rate, R_0, associated to the set L. In order to obtain optimal results randomization will be performed for 1,000 iterations. Subsequently, the set L will be ordered according to the original WfWG word-family arrangement and the effective entropy rate of the resulting set will be estimated. This value will be called r_{WfWG}. The difference, $R_0 - r_{WfWG}$, will give an estimate of the amount of absolute redundancy, D_{WfWG}, that can be attributed to the set L arranged in word families. Finally, the set L will be rearranged in an alphabetical order and its effective entropy rate will be calculated. This value will be called r_{Alpha}. The difference between the absolute rate of entropy, R_0, and the effective entropy rate, r_{Alpha}, will render an approximation to the amount of regularity, as D_{Alpha}, contained in L after imposing an alphabetical order on it.

To obtain the entropy rate value, the non-parametric estimation method developed by Kontoyiannis (1997; Kontoyiannis et al. 1998) will be implemented. This string-match method is closely related to the Lempel–Ziv–Welch compression algorithm (Welch 1984; Ziv and Lempel 1978), where H represents the average amount of entropy estimated at each position i of a string of length N[7]:

$$H = \left[\frac{1}{N}\sum_{i=2}^{N}\frac{\ell_i}{\log_2 i}\right]^{-1} \tag{2}$$

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

The magnitude of interest for the calculation of said amount at each position i is the maximum length of coincidence (maximum string length), ℓ_i. In order to determine the regularity or redundancy at position i, we must first analyze the previous segment of the string up to — but not including — i (Welch 1984), and monitor how many of the characters initials of the segment of the string starting at i have already appeared in the same order somewhere in the previous segment parsed. The value of ℓ_i is obtained by adding the unit to the length of the longest matching substring and thus meeting the following criteria: (i) it starts at position i of the string and (ii) it is not a substring of the string segment before i (Koplenig et al. 2017): "the intuitive idea behind this approach is that longer match-lengths are, on average, indicative of more redundancy in the text and, therefore, a lower mean uncertainty per character" (2017: 3). This amount of entropy contained in each character can be defined as the average amount of information in bits per character (bpc) necessary to reproduce the lemma corpus in the considered macrostructural arrangement (cf. Koplenig et al. 2017).

3. Results

Table 1 presents the results obtained for the different arrangements considered in this study. The WfWG macrostructural arrangement in word families shows a high degree of correlation between the lemma elements and, therefore, a low degree of complexity. Secondly, an alphabetic restructuring at the lexical level results in a disruption of the correlations between the elements, which leads to a higher degree of complexity in the string. Thirdly, a random restructuring removes all regularity contained in the original corpus and results in the highest value of complexity. This value is located in the upper bound for the maximum entropy of the lemma corpus at the lexical level.

Table 1: Sample of the resulting string together with the H value for each experimental setting (ES). The test sets correspond to the settings ES1 and ES2, while the control set is defined in ES3. The table includes the sample standard error (SE) and 95% confidence interval (CI) for the ES3 estimates.

Description	Sample	H Value	SE	95% CI
ES1. WfWG	hüllen hülse enthülsen einhüllen enthüllen enthüllung umhüllen verhüllen verhüllt unverhüllt human inhuman	2.31909		
ES2. Alphabetical	einhüllen enthüllen enthüllung enthülsen hüllen hülse human inhuman umhüllen unverhüllt verhüllen verhüllt	2.34744		
ES3. Random	verhüllt enthülsen hülse inhuman einhüllen verhüllen enthüllen unverhüllt umhüllen hüllen enthüllung human	2.40982	2.23733 $^{-5}$	4.38518 $^{-5}$

The difference in the value $D_{WfWG} = .09073$ as compared to $D_{Alpha} = .06238$ indicates a percent increase of 45,44% in regularity for the original word-family arrangement of the WfWG relative to the alphabetical arrangement (Figure 5).

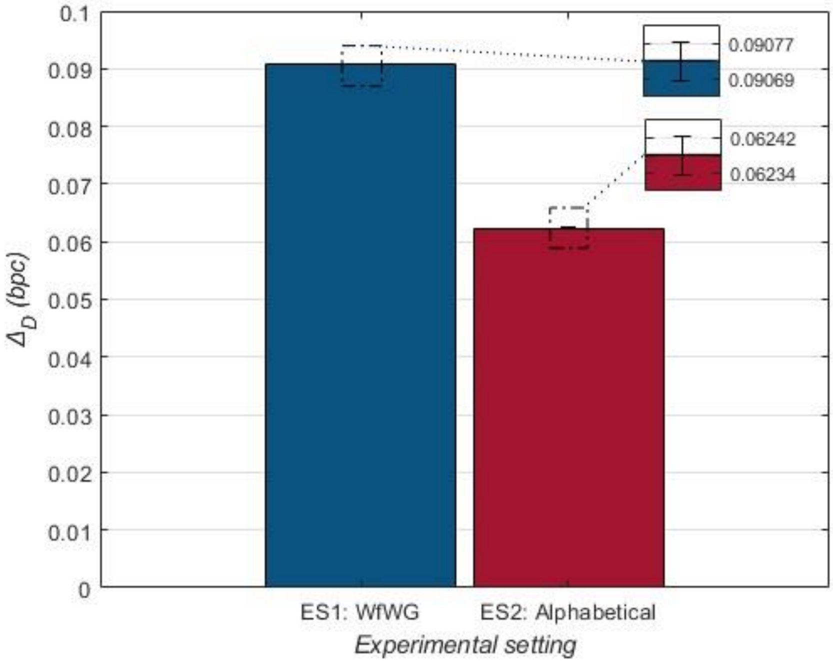


Figure 5: Representation of the increase of D_{WfWG} and D_{Alpha} in bpc according to ES1 (word families) and ES2 (alphabetical) respectively, measured against the control set established in ES3 (random). The error bars represent the 95% confidence interval for 1.000 iterations in ES3.

The increase observed in ES1 and ES2 in relation to ES3 provides a measure of the average amount of regularity or meaningful information (in bpc) gained as a function of the macrostructural arrangement of the lemma set (see Table 2). As Figure 5 shows, this difference corresponds to a percent increase of 39,12% and 26,57%, respectively.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Table 2: Amount of regularity gained by increasing the internal order of the words ($\Delta_{ESx.3}^{ESx}$) as against the amount of regularity gained by the increment in external (macrostructural) order of the words (Δ_{ES3}^{ESx})

Description	Sample	H Value	Δ_{ES3}^{ESx}	$\Delta_{ESx.3}^{ESx}$
ES1. WfWG	hüllen hülse enthülsen einhüllen enthüllen enthüllung umhüllen	2.31909	0.09073 CI: 4.38518 $^{-5}$	1.20267 CI: 4.77847^{-5}
ES3. Random	verhüllt enthülsen hülse inhuman einhüllen verhüllen enthüllen	2.40982		
ES1.3	nlhleü hsüle ehtnnüles lnhilnüee lhlteneün uelnhntglü lülmeuhn	3.52176		
ES2. Alphabetical	einhüllen enthüllen enthüllung enthülsen hüllen hülse human inhuman	2.34744	0.06238 CI: 4.38518 $^{-5}$	1.17259 CI: 4.62447^{-5}
ES3. Random	verhüllt enthülsen hülse inhuman einhüllen verhüllen enthüllen	2.40982		
ES2.3	hlünnelie hüleetnnl üetullnghn lsetnhüne ehünll lsüeh uanhm nmanhui	3.52003		

The values obtained in ES1 and ES2 in relation to ES1.3 and ES2.3 reveal the average gain of redundancy or meaningful information as a function of the intralexical order, that is, of the internal structure of words. This average gain returns values of $\Delta_{ES1.3}^{ES1}$= 1.20267 and $\Delta_{ES2.3}^{ES2}$= 1.17259, respectively. According to these data as presented in Table 3, the intralexical order in the word-family and alphabetical arrangements increases the amount of meaningful information approximately by a multiple of 13 and 17, respectively, as compared to the amount of meaningful information gained from the extralexical order of the lemma corpus.

As for the regularity increase within the alphabetic sections that make up the dictionary, Figure 6 displays the normalized absolute redundancy values (D) relative to each of the sets C_A,C_B,C_C,..., etc., except for the sets C_X and C_Y, whose content (five lemma elements in each section, with a total of 34 and 19 characters, respectively) returns an insufficient amount of data for the application of the proposed method.

In order to validate the data, the constituent elements of the test sets and the control set have been subjected to randomization in subsequent stages (Table 3). This gradual dismantling of the structures progressively suppresses any correlation and, therefore, any regularity present in the corpus.

On the other hand, as Koplenig et al. (2017) observe, the entropy rate H of any process can only be determined in the limit, that is, in strings of infinite length. In this regard, Figure 7 shows that the estimation method presented in (2) yields values that rapidly converge to the entropy source, which suggests that the obtained values yield a valid estimate of said source (ibid.).

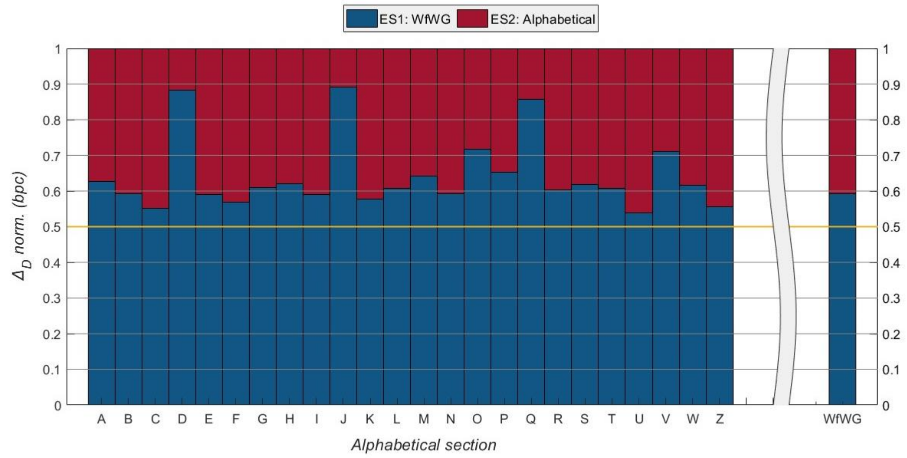


Figure 6: Normalized representation of the increase in bpc of absolute redundancy (D) for each of the alphabetical sections (except for the sets C_X and C_Y) and the whole lemma set (identified as WfWG on the rightmost side of the chart) as a function of the macrostructural arrangement per ES1 (blue) and ES2 (red) compared to the purely random arrangement in ES3. The solid orange line serves as a visual reference indicator illustrating equal values for ES1 and ES2.

Table 3: Dismantling stages for the respective experimental contexts. The third stage — represented by ES1.3, ES2.3, and ES3.3 — reproduces a version of the respective string in which the intralexical regularities have been concealed by replacing the substring containing each lexical element with another substring constructed randomly from the characters available in it (Koplenig et al. 2017). In ES3.5 the spaces have been removed from the string, subsequently the characters have been randomized and the spaces randomly re-inserted in such a way that the corpus extension remains unaltered at 27.622 instances. Results for randomized strings reflect 1,000 iterations.

		Description	Sample	H Value	SE	95% CI
ES1.	**1.1.**	**No spacing**	hüllenhülseenthülseneinhüllenenthüllenenthüllungumhüllenverhüllenverhülltunverhüllthumaninhuman	2.47437		
	1.2.	**Random spacing**	hü ll enhülseenthülseneinhü llenent hüllenenthüllu ngumhülle nve rhül lenverhül ltu nverhüllthumaninhuman	2.78207	3.6103^{-5}	7.07619^{-5}
	1.3.	**Randomized characters with original spacing**	nlhleü hsüle ehtnnüles lnhilnüee lhlteneün uelnhntglü lülmeuhn ervelühnl rlthvlüe nürtlhuvel umhna ahuninm	3.52176	2.43799^{-5}	4.77847^{-5}

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

ES2.	**2.1. No spacing**	einhüllenenthüllenenthüllungenthülsenhüllenhülsehumaninhumanumhüllenunverhülltverhüllenverhüllt	2.50761		
	2.2. Random spacing	ein hüllenenthü lle nenthüllungenthülsenhüllenhül se humaninhum anumhül l enunv erhülltverhüllenverhül lt	2.77471	3.24735^{-5}	6.36481^{-5}
	2.3. Randomized characters with original spacing	hlünnelie hüleetnnl üetullnghn lsetnhüne ehünll lsüeh uanhm nmanhui nuleühlm üvhrulnetl lülenherv ührtvlle	3.52003	2.35942^{-5}	4.62447^{-5}
ES3.	**3.1. No spacing**	verhülltenthülsenhülseinhumaneinhüllenverhüllenenthüllenunverhülltumhüllenhüllenenthüllunghuman	2.58352	2.95501^{-5}	5.79182^{-5}
	3.2. Random spacing	verhüllte nt hülsenhülse in huma neinhüllenverh üllenent h üllen unverhülltumhüllenhüllen enthüllunghuman	2.82903	3.47808^{-5}	6.81704^{-5}
	3.3. Randomized characters with original spacing	etvhrüll tlennsheü lehüs minanhu leenlihün eelvrünlh htnenülel vtulüehlnr ulhnümle leünhl lüluehtnng anhum	3.53895	2.32421^{-5}	4.55545^{-5}
	3.4. Randomized characters without spacing	onisusnpittoiieoieeaeklnnfpbolheemshrarsielgblenpefiispeetafcüeregsaenclfaiavssoesnaeelmctkrgeorinbrntzkimfue	3.56282	2.19787^{-5}	4.30784^{-5}
	3.5. Randomized characters with random spacing	sgdltaluc tineen ainc rgzatzrsrneb ihbss ntsviikk nnhueur nieguh wbrsedcgmeenkrat ssnakhtek iarseeecoe äeheudzane	3.571009	2.06591^{-5}	4.04917^{-5}

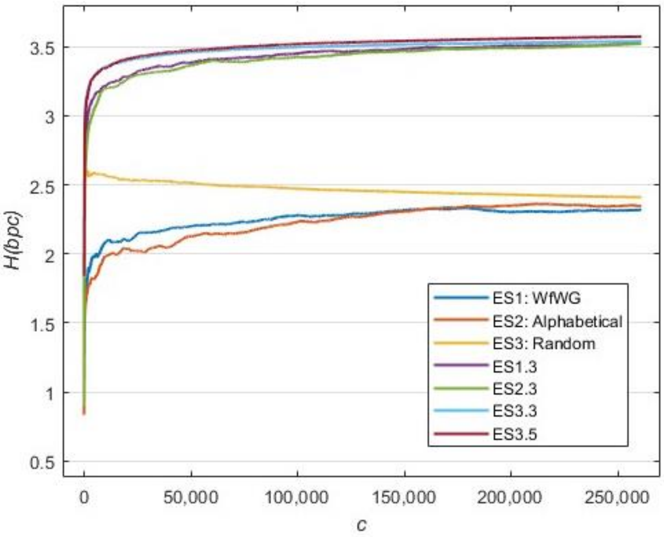


Figure 7: Entropy rate H in bpc as a function of incorporated lemmata expressed as c. The results attest that a small amount of data is enough to demonstrate a convergence towards the entropy source (Koplenig et al. 2017). All three experimental settings ESx.3 (randomized characters with original spacing) deal with the internal order

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

of words in their respective macrostructural arrangement. In ES3.5 (randomized characters with random spacing), after removing the spaces from the string, the characters have been randomized and the spaces randomly re-inserted into the string in such a way that the total amount of 'words' remains unaltered at 27.622 instances.

4. Discussion

The specific objectives of this study have focused on (O1) determining whether the macrostructural arrangement has an impact on the AC value of the WfWG corpus, (O2) calculating the difference between the corpus macrostructural arrangements — in word families and alphabetical — with regard to their AC values, as well as (O3) identifying the macrostructural arrangement that delivers the less complex version of said corpus. For this purpose, the Minimum Description Length principle has been applied to estimate the corpus AC value based on its compressibility. Among the findings, with reference to O1 the fundamentals in the data show that the macrostructural arrangement influences the AC value of the WfWG corpus. Concerning O2, the data also reveal that alterations in the macrostructure towards more ordered arrangements significantly decrease the AC value of the corpus. However, the most significant finding of this study, in relation to O3, reveals that the WfWG corpus in a word-family arrangement has a lower AC value in comparison to the value associated with the same corpus arranged in alphabetical order (see Table 1). As observed from Figure 5, said relative minimum value represents a significant increase in regularity for the word-family distribution, approximately doubling the regularity gain of the alphabetical layout relative to a purely random arrangement. On the other hand, the estimates for each alphabetic section of the dictionary displayed in Figure 6 manifest that, as for the entire set, the arrangement in word families contributes to lower AC values. The fluctuation in the relative values across the alphabetic sections points to a variation in the numerical proportion of word families in relation to the number of those lemmata that do not appear attached to any of the listed word families. Furthermore, in light of the validation procedures applied, the consistency of the results suggests that the Minimum Description Length principle renders a valid method to estimate the complexity associated with the corpus macrostructural arrangement. In more general terms, an AIT-related linguistic interpretation of the results allows us to argue that analogy is language's algorithmically simpler and, therefore, more efficient operation mode, driving it — as a self-regulating natural system — towards a more ordered configuration distant from entropic degradation (Devine 2020). In this sense, the present study inscribes itself in the area of research dedicated to the widely analyzed and documented phenomenon of analogy as a fundamental cognitive strategy in language processing.

Notwithstanding the coherence of the results with respect to the guiding principles of our proposal, this approach reveals certain limitations. An inherent

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

drawback resides in the restriction of the corpus to a set of lexical items whose superior demarcation is the word. This excludes the consideration of (to a greater or lesser degree) lexicalized syntactic constructions, for example, constructions with a functional verb, or *Funktionsverbgefüge*, in which the grammatical function of the construct is expressed by the verb. Additionally, and due to the purely numerical nature of this approach, the results do not allow us to present explicit-declarative statements about the grammar — which as a description of minimum length (Kornai 2008) governs the corpus — nor about the nature of the structures or patterns affected by the compression (both at the intralexical and interlexical level). An observation that, on the other hand, suggests that grammar acquisition through inductive inference is related to the formation of (primarily) implicit knowledge in procedural memory (DeKeyser 2015; Paradis 2009; Ullman 2016). In addition, constraining the arrangement criteria of the test sets to both word-family and alphabetical principles leaves semantic-oriented arrangement criteria unconsidered. In this regard, since cognitive-semantic criteria overlook the formal component of the linguistic sign, the macrostructural arrangements conforming to purely semantic precepts shall be deemed random per the approach adopted in this research, with their AC estimates expected to surpass the AC numbers for word-family and alphabetical arrangements, leaning towards maximum entropy values. However, the most important limitation of this study lies in the reduction of the object of analysis to a single language with distinct typological characteristics. If, as the study by Koplenig et al. (2017) suggests, the entropy rate of any language lemma corpus is determined by morphological factors, we believe that future studies including the morphological typology as an independent variable can introduce additional evidence that would validate the approach pursued in this study.

In contrast to the prevalent approaches in monolingual learner lexicography based on theoretical frameworks of cognitive semantics, our proposal introduces a psycho-cognitive approach based on a quantifiable, irreducible, and theory-neutral notion of complexity and, by extension, of learnability. In this regard, although the macrostructural arrangement in word families has been widely applied and referred to in the German lexicographic tradition as a method to facilitate language learning (by exposing the internal mechanisms of lexical production), these results suggest that the methodology adopted in this study would enable, for the first time, a theory-neutral quantitative evaluation of the didactic nature of this lexicographic practice.

On a separate note, this proposal based on the complexity-learnability binomial also supports the idea of implementing a differentiated approach in the area of digital lexicography (Bothma 2011, 2017; Bothma et al. 2016) whose interest would be focused, beyond optimal searching, in the navigational (browsing) processing of the lemma set. According to the psycho-cognitive and organizational foundations of Neuroergonomics (Lapeyre et al. 2011; Li and Klippel 2016; Montello 2005; Montello and Sas 2006), navigation in less complex (irrespective of the formal definition of complexity applied) and thus more regular environments promotes spatial awareness and the creation of more complete and precise cognitive maps that facilitate orientation and, thereby, the understanding of the environment or search space considered, in our case, the lemma corpus.

Finally, we consider it necessary to recapitulate that the current study does not address the quantitative determination of the corpus learnability. In this respect, an empirical analysis that would provide a quantitative measure of the corpus learnability as a function of the AC value associated to its macrostructural arrangement represents the major research objective of future studies within the framework of this project.

5. Conclusions

The general objective of this study is to promote language acquisition in monolingual learner's dictionary users from a psycho-cognitive perspective. In our proposal, we view the human cognition essentially as a learner sensitive to fluctuations in the environmental complexity in such a way that our learning efficiency increases in less complex environments (Kempe et al. 2015; Zenil and Gauvrit 2017). Against this background, our methodological approach — the Minimum Description Length principle — allowed us to obtain a quantitative estimation of the algorithmically bound complexity (AC) attributed to the WfWG lemma corpus according to its macrostructural arrangement. The resulting data indicate that, compared to the alphabetical layout, the arrangement in word families provides a more ordered, less complex, and, by extension, more learnable version of the lemma corpus. These results open a door to future studies with the aim to determine the lemma corpus variation in learnability as a function of the AC value derived from its macrostructural arrangement. We hope that this psycho-cognitive approach based upon the principles and practical methods of AIT may well be useful in implementing macrostructural designs in monolingual learner lexicography (both paper and digital) which would reinforce language acquisition in the dictionary user.

Acknowledgements

Our deep appreciation and gratitude are due to those whose inestimable peer review and editorial guidance vastly enhanced our original draft. Any remaining errors, omissions, and limitations are entirely ours.

Endnotes

1. De Schryver (2013) points out that in computational lexicography, analogous to printed lexicography, the macrostructural design refers not only to the dictionary as an ordered arrangement of the lemma corpus but mainly to the set of interlexical relationships that are configured around parameters such as grammatical category, morphology, valence, semantic features, etc.
2. The macrostructural treatment in machine-readable dictionaries of this particular type tends to reflect a methodical and consistent digitalization of the same theoretical principles and, beyond

the formal qualities of the new medium (menus, hypertexts, multimedia, etc.), they do not differ substantially from their printed counterparts (Chen 2012; Dziemianko 2017; Kobayashi 2007).

3. This approach represents a concretion in the area of monolingual learner lexicography of the "simplicity principle" applied in the field of SLA (Chater and Vitányi 2007; Chater et al. 2015). According to this principle, "the learner has sufficient data to learn successfully from positive evidence if it favors the simplest encoding of the linguistic input [emphasis in the original]" (Hsu et al. 2013: 35).

4. A didactically motivated approach to the word-family arrangement in German lexicographic tradition dates back as far as 1700 with the publication of *Das herrlich grosse teutsch–italiänische Dictionarium* by Matthias Kramer (cf. Haß-Zumkehr 2012: 81-88). Kramer relates his decision in favor of a word-family arrangement to the practical didactic requirements for the production of new words, that is, to the apprehension of the language internal (grammatical) mechanisms (ibid.: 84-85). Hence, the macrostructural layout in word families of his dictionary is essentially justified by its didactic purpose, an attribute that, in Kramer's words, prevails over its functionality as a reference work (Wiegand 1998: 657).

5. Consistent with Cruse's definition, a "lexical unit" designates "the union of a lexical form and a single sense" (1986: 77) as opposed to a "lexeme", a term that refers to a cluster of lexical units: "lexemes, on the other hand, are the items listed in the lexicon, or 'ideal dictionary', of a language" (Cruse 1986: 49). The lexeme congregates a group of lexical units sharing a common root and therefore, maintaining a certain relationship, both in their phonetic composition and in their meaning (Bergenholtz and Tarp 1995; Umbreit 2011): "[A] dictionary contains (among other things) an alphabetical list of the lexemes of a language. We shall characterise a lexeme as a family of lexical units" (Cruse 1986: 76). A "lexical item", on the other hand, designates "any word, abbreviation, partial word, or phrase which can figure in a dictionary (often as the headword of an entry) as the 'target' of some form of lexicographic description, most commonly a definition or a translation" (Atkins and Rundell 2008: 163).

6. Augst points here to the primary, and often fuzzy distinction between analogy and rule. In this regard, Kiparsky (1975, in Derwing and Skousen 1989: 56) argues that it is problematic to draw a clear, rigorous, and unequivocal boundary between both, since "at the point at which [...] analogies begin to make the right generalizations, they are indistinguishable from rules." Along the same lines, Haspelmath (2002: 103) considers morphological analogy and regularity as "really one and the same thing", while in the words of Krott (2009: 118) rules can be qualified "as extreme case of analogy".

7. Koplenig et al. (2017) remark that, since this is an estimate of the theoretical (and unobservable) value, the correct mathematical notation corresponds to the hat notation $\hat{H}$. In this regard, we adhere to the authors' motivation and, for the sake of simplicity, we adopt the plain notation H.

References

A. Dictionaries

Augst, Gerhard. 2009. *Wortfamilienwörterbuch der deutschen Gegenwartssprache*. Tübingen: Niemeyer. https://doi.org/10.1515/9783484971332

Benson, Morton, Evelyn Benson and Robert Ilson. 2009. *The BBI Combinatory Dictionary of English. Your Guide to Collocations and Grammar*. Amsterdam: John Benjamins.

Bosque, Ignacio (Ed.). 2006. *Diccionario combinatorio práctico del español contemporáneo. Las palabras in su contexto.* Madrid: SM.

Casares, Julio. 2013. *Diccionario ideológico de la lengua española. Desde la idea a la palabra, desde la palabra a la idea.* Barcelona: Gustavo Gili.

Davau, Maurice, Marcel Cohen and Maurice Lallemand. 1984. *Dictionnaire du français vivant*. Paris: Bordas.

Dornseiff, Franz. 2020. *Der deutsche Wortschatz nach Sachgruppen*. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110457742

Häcki Buhofer, Annelies, Marcel Dräger, Stefanie Meier and Tobias Roth. 2014. *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke.

Kirkpatrick, Elizabeth M. (Ed.). 1983. *Chambers Universal Learners' Dictionary*. Edinburgh: Chambers.

Mel'čuk, Igor A., Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha and Alain Polguère. 1999. *Dictionnaire explicatif et combinatoire du français contemporain*. Montreal: Montreal University Press.

Rey-Debove, Josette and Alain Rey. 2009. *Le nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française.* Paris: Dictionnaires Le Robert.

Simone, Raffaele. 2010. *Grande Dizionario Analogico della Lingua Italiana.* Turin: UTET.

B. Other Literature

Atkins, B.T. Sue and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Augst, Gerhard. 1975. *Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache.* Tübingen: Narr.

Augst, Gerhard. 1992. Das lexikologische Phänomen der Wortfamilie in alphabetisch-semasiologischen Wörterbüchern. *Zeitschrift für germanistische Linguistik* 20(1): 24-36. https://doi.org/10.1515/zfgl.1992.20.1.24

Bergenholtz, Henning and Sven Tarp (Eds.). 1995. *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries.* Amsterdam: John Benjamins.

Bothma, Theo J.D. 2011. Filtering and Adapting Data and Information in an Online Environment in Response to User Needs. Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography. The Internet, Digital Initiatives and Lexicography*: 71-102. London: Bloomsbury Academic.

Bothma, Theo J.D. 2017. Lexicography and Information Science. Fuertes-Olivera, P.A. (Ed.). 2017. *The Routledge Handbook of Lexicography*: 197-216. London: Routledge. https://doi.org/10.4324/9781315104942-14

Bothma, Theo J.D., Rufus H. Gouws and Danie J. Prinsloo. 2016. The Role of e-Lexicography in the Confirmation of Lexicography as an Independent and Multidisciplinary Field. Margalitadze, T. and G. Meladze (Eds.). 2016. *Proceedings of the 17th EURALEX International Congress. Lexicography and Linguistic Diversity, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia, 6–10 September 2016*: 109-116. Tbilisi: Ivane Javakhishvili Tbilisi State University.

Chaitin, Gregory J. 1969. On the Length of Programs for Computing Finite Binary Sequences. Statistical Considerations. *Journal of the ACM* 16(1): 145-159. https://doi.org/10.1145/321495.321506

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Chaitin, Gregory J. 2004. *Algorithmic Information Theory*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511608858

Chaitin, Gregory J. 2006. The Limits of Reason. *Scientific American* 294(3): 74-81.

Chater, Nick, Alexander Clark, John A. Goldsmith and Amy Perfors. 2015. *Empiricism and Language Learnability*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198734260.001.0001

Chater, Nick and Paul Vitányi. 2007. 'Ideal Learning' of Natural Language. Positive Results about Learning from Positive Evidence. *Journal of Mathematical Psychology* 51(3): 135-163. https://doi.org/10.1016/j.jmp.2006.10.002

Chen, Yuzhen. 2012. Dictionary Use and Vocabulary Learning in the Context of Reading. *International Journal of Lexicography* 25(2): 216-247. https://doi.org/10.1093/ijl/ecr031

Clark, Alexander and Shalom Lappin. 2013. Complexity in Language Acquisition. *Topics in Cognitive Science* 5(1): 89-110. https://doi.org/10.1111/tops.12001

Cruse, David A. 1986. *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

David, Ofir, Shay Moran and Amir Yehudayoff. 2016. Supervised Learning through the Lens of Compression. Lee, D. et al. (Eds.). 2016. *Advances in Neural Information Processing Systems 29. 30th Conference on Neural Information Processing Systems, Barcelona, 5–10 December 2016:* 2784-2792.

DeKeyser, Robert. 2015. Skill Acquisition Theory. Van Patten B. and J. Williams (Eds.). 2015. *Theories in Second Language Acquisition*: 94-112. Second edition. New York: Routledge.

Derwing, Bruce L. and Royal Skousen. 1989. Morphology in the Mental Lexicon. A New Look at Analogy. Booij, G. and J. van Marle (Eds.). 1989. *Yearbook of Morphology 1989. Volume 2*: 55-71. Dordrecht: Kluwer Academics. https://doi.org/10.1515/9783112420560-005

De Schryver, Gilles-Maurice. 2013. Tools to Support the Design of a Macrostructure. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography:* 1384-1395. Berlin/Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110238136.1384

Devine, Sean. 2020. *Algorithmic Information Theory for Physicists and Natural Scientists*. Bristol: IOP Publishing. https://doi.org/10.1088/978-0-7503-2640-7

Dong, Zhendong and Qiang Dong. 2006. *HowNet and the Computation of Meaning*. Singapore: World Scientific. https://doi.org/10.1142/5935

Dziemianko, Anna. 2017. Dictionary Form in Decoding, Encoding and Retention: Further Insights. *ReCALL* 29(3): 335-356. https://doi.org/10.1017/S0958344017000131

Engelberg, Stefan and Lothar Lemnitzer. 2009. *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Evans, Vyvyan. 2009. *How Words Mean. Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199234660.001.0001

Fillmore, Charles J. 1982. Frame Semantics. *Linguistics in the Morning Calm. Selected Papers from SICOL-1981*: 111-137. Seoul: Hanshin.

Fillmore, Charles J. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica* 6(2): 222-254.

Fillmore, Charles J., Christopher R. Johnson and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography* 16(3): 235-250. https://doi.org/10.1093/ijl/16.3.235

Fuertes-Olivera, Pedro A. and Sven Tarp. 2011. Lexicography for the Third Millennium. Cognitive-oriented Specialised Dictionaries for Learners. *Ibérica* 21: 141-161. https://www.revistaiberica.org/index.php/iberica/article/view/332

Fulop, Sean A. and Nick Chater. 2013. Learnability Theory. *WIREs Cognitive Science* 4(3): 299-306. https://doi.org/10.1002/wcs.1228

Geeraerts, Dirk. 2001. The Definitional Practice of Dictionaries and the Cognitive Semantic Conception of Polysemy. *Lexicographica* 17: 6-21. https://doi.org/10.1515/9783110244212.6

Gouws, Rufus H. 2003. Types of Articles, their Structure and Different Types of Lemmata. Van Sterkenburg, P. (Ed.). 2003. *A Practical Guide to Lexicography*: 34-43. Amsterdam: John Benjamins.

Grünwald, Peter. 2005. Introducing the Minimum Description Length Principle. Grünwald, P., J.I. Myung and M.A. Pitt (Eds.). 2005. *Advances in Minimum Description Length. Theory and Applications*: 3-22. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/1114.003.0004

Grünwald, Peter. 2007. *The Minimum Description Length Principle*. Cambridge, MA: The MIT Press.

Grünwald, Peter and Paul Vitányi. 2008. Algorithmic Information Theory. Adriaans, P. and J. van Benthem (Eds.). 2008. *Philosophy of Information*: 289-325. Amsterdam: Elsevier. https://doi.org/10.48550/arXiv.0809.2754

Haspelmath, Martin. 2002. *Understanding Morphology*. London: Hodder Education.

Haß-Zumkehr, Ulrike. 2012. *Deutsche Wörterbücher: Brennpunkt von Sprach- und Kulturgeschichte*. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110849189

Hausmann, Franz Josef and Herbert Ernst Wiegand. 1989. Component Parts and Structures of General Monolingual Dictionaries. A Survey. Hausmann, F.J., O. Reichmann, H.E. Wiegand and L. Zgusta (Eds.). 1989. *Wörterbücher. Ein internationales Handbuch zur Lexikographie. Volume 1:* 328-360. Berlin: De Gruyter.

Hsu, Anne S., Nick Chater and Paul Vitányi. 2013. Language Learning from Positive Evidence, Reconsidered. A Simplicity-based Approach. *Topics in Cognitive Science* 5(1): 35-55. https://doi.org/10.1111/tops.12005

Kempe, Vera, Nicolas Gauvrit and Douglas Forsyth. 2015. Structure Emerges Faster during Cultural Transmission in Children than in Adults. *Cognition* 136: 247-254. https://doi.org/10.1016/j.cognition.2014.11.038

Kiparsky, Paul. 1975. What Are Phonological Theories About? Cohen, D. and J.R. Wirth (Eds.). 1975. *Testing Linguistic Hypotheses*: 187-209. Washington: Wiley.

Kobayashi, Chiho. 2007. Comparing Electronic and Printed Dictionaries: Their Effects on Lexical Processing Strategy Use, Word Retention, and Reading Comprehension. Bradford-Watts, K. (Ed.). 2007. *JALT 2006 Conference Proceedings*: 657-671. Tokyo: Japan Association of Language Teaching.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Kolmogorov, Andrei N. 1963. On Tables of Random Numbers. *The Indian Journal of Statistics*. Series A 25(4): 369-376.

Kontoyiannis, Ioannis. 1997. The Complexity and Entropy of Literary Styles. *NSF Technical Report* 97: 1-15.
https://purl.stanford.edu/nw057vj8228

Kontoyiannis, Ioannis, Paul H. Algoet, Yuri M. Suhov and Abraham J. Wyner. 1998. Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text. *IEEE Transactions on Information Theory* 44(3): 1319-1327.
https://doi.org/10.1109/18.669425

Koplenig, Alexander, Peter Meyer, Sascha Wolfer, Carolin Müller-Spitzer and Kenny Smith. 2017. The Statistical Trade-off between Word Order and Word Structure. Large-scale Evidence for the Principle of Least Effort. *PLoS ONE* 12(3): 1-25.
https://doi:10.1371/journal.pone.0173614

Kornai, András. 2008. *Mathematical Linguistics*. London: Springer.
https://doi.org/10.1007/978-1-84628-986-6

Kövecses, Zoltán and Szilvia Csábi. 2014. Lexicography and Cognitive Linguistics. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 27(1): 118-139.
https://doi.org/10.1075/resla.27.1.05kov

Kremer, Gerhard, Andrea Abel and Marco Baroni. 2008. Cognitively Salient Relations for Multilingual Lexicography. Zock, M. and C.-R. Huang (Eds.). 2008. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), Manchester, 24 August 2008:* 94-101. Coling 2008 Organizing Committee. https://aclanthology.org/W08-1913

Krott, Andrea. 2009. The Role of Analogy for Compound Words. Blevins, J.P. and J. Blevins (Eds.). 2009. *Analogy in Grammar. Form and Acquisition*: 118-136. Oxford: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780199547548.003.0006

Lakoff, George. 1993. The Contemporary Theory of Metaphor. Ortony, A. (Ed.). 1993. *Metaphor and Thought*: 202-251. Cambridge, UK: Cambridge University Press.
https://doi.org/10.1017/CBO9781139173865.013

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lapeyre, Brigitte, Sylvain Hourlier, Xavier Servantie, Bernard N'Kaoua and Hélène Sauzéon. 2011. Using the Landmark–Route–Survey Framework to Evaluate Spatial Knowledge Obtained from Synthetic Vision Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53(6): 647-661.
https://doi.org/10.1177/0018720811421171

Li, Rui and Alexander Klippel. 2016. Wayfinding Behaviors in Complex Buildings. *Environment and Behavior* 48(3): 482-510.
https://doi.org/10.1177/0013916514550243

Maguire, Phil, Oisín Mulhall, Rebecca Maguire and Jessica Taylor. 2015. Compressionism: A Theory of Mind Based on Data Compression. Airenti, G., B.G. Bara and G. Sandini (Eds.). 2015. *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science / 4th European Conference on Cognitive Science / 11th International Conference on Cognitive Science, Torino, Italy, 25–27 September 2015*: 294-299. CEUR Workshop Proceedings.

Martínez de Sousa, José. 2009. *Manual básico de Lexicografía*. Gijón: Ediciones Trea.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. 1990. Introduction to WordNet. An On-line Lexical Database. *International Journal of Lexicography* 3(4): 235-244.
https://doi.org/10.1093/ijl/3.4.235

Montello, Daniel R. 2005. Navigation. Shah, P. and A. Miyake (Eds.). 2005. *The Cambridge Handbook of Visuospatial Thinking*: 257-294. Cambridge, UK: Cambridge University Press.
https://doi.org/10.1017/CBO9780511610448.008

Montello, Daniel R. and Corina Sas. 2006. Human Factors of Wayfinding in Navigation. Karwowski, W. (Ed.). 2006. *International Encyclopedia of Ergonomics and Human Factors*: 2003-2008. Boca Raton, FL: CRC Press.
http://dx.doi.org/10.1201/9780849375477.ch394

Ostermann, Carolin. 2015. *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics*. Berlin/Boston: De Gruyter.
https://doi.org/10.1515/9783110424164

Paradis, Michel. 2009. *Declarative and Procedural Determinants of Second Languages.* Philadelphia: John Benjamins.

Richardson, Stephen D., William B. Dolan and Lucy Vanderwende. 1998. MindNet: Acquiring and Structuring Semantic Information from Text. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL–Coling), Montreal, Quebec, Canada, August 1998. Volume 2:* 1098-1102. Montreal: Association for Computational Linguistics.
https://dl.acm.org/doi/10.3115/980691.980749

Rissanen, Jorma. 1978. Modeling by Shortest Data Description. *Automatica* 14(5): 465-471.
https://doi.org/10.1016/0005-1098(78)90005-5

Solomonoff, Ray J. 1964a. A Formal Theory of Inductive Inference. Part I. *Information and Control* 7(1): 1-22.
https://doi.org/10.1016/S0019-9958(64)90223-2

Solomonoff, Ray J. 1964b. A Formal Theory of Inductive Inference. Part II. *Information and Control* 7(2): 224-254.
https://doi.org/10.1016/S0019-9958(64)90131-7

Tyler, Andrea and Vyvyan Evans. 2004. Applying Cognitive Linguistics to Pedagogical Grammar: The Case of *Over*. Achard, M. and S. Niemeier (Eds.). 2004. *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching*: 257-280. Berlin/New York: De Gruyter Mouton.
https://doi.org/10.1515/9783110199857.257

Ullman, Michael T. 2016. The Declarative/Procedural Model: A Neurobiological Model of Language Learning, Knowledge, and Use. Hickok, G. and S.L. Small (Eds.). 2016. *Neurobiology of Language*: 953-968. London: Academy Press.
https://doi.org/10.1016/B978-0-12-407794-2.00076-6

Umbreit, Birgit. 2011. Motivational Networks: An Empirically Supported Cognitive Phenomenon. Panther, K.-U. and G. Radden (Eds.). 2011. *Motivation in Grammar and the Lexicon*: 269-286. Amsterdam: John Benjamins.
https://doi.org/10.1075/hcp.27.17umb

Welch, Terry A. 1984. A Technique for High-performance Data Compression. *Computer* 17(6): 8-19.
https://doi.org/10.1109/MC.1984.1659158

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Wiegand, Herbert Ernst. 1989. Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch. Alphabetische Anordnungsformen und ihre Probleme. Hausmann, F.J., O. Reichmann, H.E. Wiegand and L. Zgusta (Eds.). 1989. *Wörterbücher. Ein internationales Handbuch zur Lexikographie. Volume 1:* 371-409. Berlin: De Gruyter.
http://dx.doi.org/10.1515/9783110095852.1.4.328

Wiegand, Herbert Ernst. 1998. Historische Lexikographie. Besch, W., A. Betten, O. Reichmann and S. Sonderegger (Eds.). 1998. *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Volume 1:* 643-713. Berlin/New York: De Gruyter.

Zenil, Hector and Nicolas Gauvrit. 2017. Algorithmic Cognition and the Computational Nature of the Mind. Meyers, R.A. (Ed.). 2017. *Encyclopedia of Complexity and Systems Science*: 1-9. New York: Springer.
https://doi.org/10.1007/978-3-642-27737-5_707-2

Ziv, Jacob and Abraham Lempel. 1978. Compression of Individual Sequences via Variable-rate Coding. *IEEE Transactions on Information Theory* 24(5): 530-536.
https://doi.org/10.1109/TIT.1978.1055934