Quantitative Insights into Syllabic Structures

by

Peter Zörnig
Kamil Stachowski
Anna Rácová
Yunhua Qu
Michal Místecký
Kuizi Ma
Mihaiela Lupea
Emmerich Kelih
Volker Gröller
Hanna Gnatchuk
Alfiya Galieva
Sergey Andreev
Gabriel Altmann

2019
RAM-Verlag

Studies in Quantitative Linguistics

Editors

Andreev, Sergey	(smol.an@mail.ru)
Emmerich Kelih	(emmerich.kelih@univie.ac.at)
Reinhard Köhler	(koehler@uni-trier.de)
Haitao Liu	(htliu@163.com)
Ján Mačutek	(jmacutek@yahoo.com)
Místecký, Michal	(MMistecky@seznam.cz)
Eric S. Wheeler	(wheeler@ericwheeler.ca)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro.* 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative Linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3.* 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4.* 2014, VIII+148 pp.
15. Best, K.-H., Kelih, E. (eds.), *Entlehnungen und Fremdwörter: Quantitative Aspekte.* 2014. VI + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, G. *Unified Modeling of Length in Language.* 2014, VIII + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical Approaches to Text and Language Analysis dedicated to Luděk Hřebíček on the occasion of his 80th birthday.* 2014. VI + 231 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA Quantitative Index Text Analyzer*. 2014, VII + 106 pp.

19. K.-H. Best, *Studien zur Geschichte der Quantitativen Linguistik.* 2015. III + 158 pp.
20. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, P. Mohanty, E. Kelih, R. Chen, G. Altmann, *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* 2015. IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015. III+146 pp.
22. P. Zörnig, K. Stachowski, I.-I. Popescu, T. Mosavi Miangah, R. Chen, G. Altmann, *Positional Occurrences in Texts: Weighted Consensus Strings.* 2015. II+178 pp.
23. E. Kelih, R. Knight, J. Mačutek, A.Wilson (eds.), *Issues in Quantitative Linguistics Vol. 4*. 2016. III + 231 pp.
24. J. Léon, S. Loiseau (eds.), *History of quantitative linguistics in France*. 2016. II + 232 pp.
25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, III+171 pp.
26. M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme,* 2017, IV + 134 pp.
27. G. Altmann, *Unified Modeling of Diversification in Language*. 2018, VIII+119 pp.
28. E. Kelih, G. Altmann, *Problems in Quantitative Linguistics, Vol. 6.* 2018, IX+118 pp.
29. S. Andreev, M. Místecký, G. Altmann, *Sonnets: Quantitative Inquiries.* 2018, VI + 129 pp
30. P. Zörnig, K.Stachowski, A. Rácová, Y. Qu, M. Místecký, K. Ma, M. Lupea, E. Kelih, V. Gröller, H. Gnatchuk, A. Galieva, S. Andreev, G. Altmann, *Quantitative Insights into Syllabic Structures.* 2019, VI + 133

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
Germany
RAM-Verlag@t-online.de
http://ram-verlag.eu

ISBN: 978-3-942303-88-0

Contents

1. Introduction

1.1 Basic syllable models

The position of the syllable in linguistics is not undisputed. Mostly, the missing transparent and clear definition of this unit seems to be the major argument for its banishment from the linguistic discussion. This position is well reflected in Kohler (1966: 207), where he critically emphasizes:

> The syllable is very often regarded as a substantive universal in phonology; but it can be demonstrated that the syllable is either an UNNECESSARY concept, because the division of the speech chain into such units is known for other reasons, or an IMPOSSIBLE one, as any division would be arbitrary, or even a HARMFUL one, because it clashes with grammatical formatives. If the syllable has any real status in phonology, its boundaries must be discernible.

This assessment has to be seen in the light of the linguistic discussion of the 1960s, regarding the priority of phonetic or phonological approaches. Moreover, taking into consideration recently dominating phonological theories (optimality theory, lexical and prosodic phonology, natural phonology, and in general "preference"-based approaches), it appears that there is no lack of suggestions regarding a proper definition of the syllable, and in particular a linguistically grounded syllable division, e.g. the determination of syllable boundaries. The fact that the syllable is in the ongoing focus of linguistics goes hand in hand with the elaboration of different models of it, which will be briefly presented in the following.

One has to begin with the most simple syllable (σ) model, consisting of three constituents. The most important one is the syllable nucleus, characterized by a high degree of sonority, and thus usually equalling a vocalic segment. Before the nucleus, the syllable head is located, which is also termed as syllable onset or onset only. After the nucleus, the syllable coda is located (cf. Fig. 1.1, based on van der Hulst/Ritter 1999: 38 and Fudge 1987: 3).

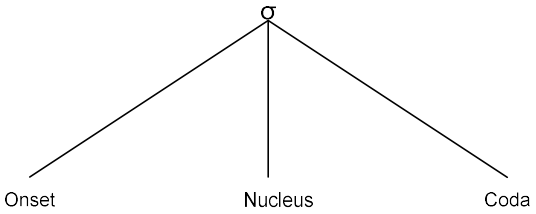

Fig. 1.1. Syllable model: onset – nucleus – coda

This tripartite model is common, both in (older) structuralistic references and in newer approaches, like optimality theory (cf. Archangeli 1997, Hammond 1997: 36, Kager 1999: 91). Although the model lacks a further hierarchy, it is nevertheless due to its simplicity regarded as a basic model in syllable phonology.

An alternative view on the syllable is achieved by merging the nucleus and coda into a common constituent, which is usually called *rhyme*, or *rime* (cf. van der Hulst/Ritter 1999: 22, Fudge 1987: 360). In phonology, several arguments have been raised in favour of the onset-rhyme model, which Fudge (1987: 376) sees as "[…] the best model for the syllable […]." First, it is well known that phonotactic restrictions almost apply for the rime. Secondly, it appears that in language games and slips of the tongue mostly the rime is affected and not only some sub-constituents. Thirdly, there is empirical evidence (cf. Treiman/Kessler 1995) for an intuitive segmentation of the syllable by native speakers into the onset and the rime, which favours the psycho-linguistic reality of these units. The bipartite model is also of interest in case of considering the accentual and prosodic structures, where one can distinguish heavy and light syllables (cf. Vater 1992: 125–126).

A further alternative of a bipartite model is a body-tail model (cf. van der Hulst/Ritter 1999: 22), where the onset and the nucleus form the syllable body, followed by the coda. However, this model is much less discussed and "applied" than the previously mentioned ones.

A rather minimalistic approach to the syllable (cf. Clements/Keyser 1983 and Hyman 1985) is its reduction to the constituting consonants (C) and vowels (V); this is usually referred to as the skeleton tier.

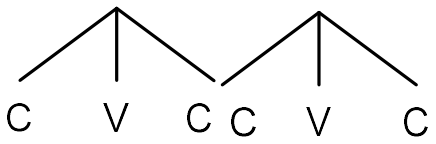

Fig. 1.2. CV structure of syllables (Clements/Keyser 1983: 8)

As can be seen from Fig. 1.2, this kind of representation easily allows the addition of a further specification of the vocalic nucleus, to which length and other prosodic features can be added. Therefore, this model is popular in particular for the description of quantity-sensitive languages.

Finally, a further syllable model is the mora or the moraic syllable (sometimes also called rime). The mora is a phonological measurement unit in a short syllable, consisting of one short vowel and maximally one consonant; bimoraic syllables are consisting of a syllable with a long vowel, or a short vowel and two or more consonants (cf. Fig. 1.3). The moraic syllable is therefore directly related (cf. van der Hulst/Ritter 1999: 28) with the concept of syllable weight, where the vowel quantity and the vowel length play immanent roles.

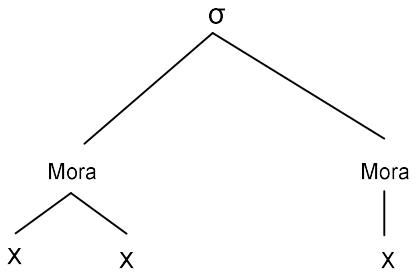

Fig. 1.3. Moraic syllable (Clements/Keyser 1983)

In phonology, as can be drawn from the above brief overview, several models of the syllable are indeed at disposal. There can be no definitely "adequate" model, since the

relevance of a model is directly related to the particular linguistic problem analysed. Moreover, in addition to several suggestions regarding the definition of the syllable, the syllabification – e.g., the determination of the syllable boundaries – is much more challenging. Some basic aspects of this problem will be presented in the next section.

1.2 The syllable: domain and processes

The syllable is a phonological, phonetic, and prosodic unit. Moreover, it is the domain of phonological and phonetic processes, such as, for instance, aspiration, regressive/ progressive assimilation, pharyngealization, etc. According to Donegan/Stampe (1979: 142ff.), mainly fortition processes (strengthening processes) – which intensify the salient features of individual segments and/or their contrast (dissimilation, diphthongization, syllabification, and epenthesis) – can be distinguished from lenition processes (assimilation, monophthongization, desyllabification, reduction, deletion), making segments and sequences of segments easier to pronounce. For both fortition and lenition processes, the syllable appears to be a proper framework of description and analysis.

A further important domain of the syllable are prosodic characteristic of languages; in particular, it is believed that the syllable is the bearer of the tone, the accent, and/or the stress. Moreover, for the study of prosody and intonation, the syllable usually seems to be the proper reference unit (for further details on prosodic and metric phonology, see Hayes 1995, Hyman 1985, and Itô 1988). A particularly important role is played by the syllable in phonotactics and phoneme distribution (cf. Blevins 1995, van der Hulst/Ritter 1999: 20f., Greenberg 1978, Sigurd 1955, 1965, Algeo 1978, Basbøll 1999, Hall 2000: 230, Vestergaard 1967, Ewen/van der Hulst 2001: 123ff., O'Connor/Trim 1953, Haugen 1956, Archangeli 1997: 8ff.). In this kind of research, the focus is laid on the compatibility of particular phonemes and positional constraints of phonemes. In this research, the syllable is usually considered to be a proper reference unit; however, units like morphems, word forms, etc., can also come into play.

In addition to being the core unit of phonetics and phonology, the syllable is referred to in many other linguistics domains, too. Among others, the syllable is relevant for psycholinguistics, including language games with some interchange of phonological segments, slips of the tongue, reversing of phonemes in a word or syllable. In general, the syllable can also be considered a basic unit of language processing, being part of a phonetic syllable or mental syllable lexicon, as supposed by Levelt (1992), Levelt/Wheeldon (1994), Schiller et al. (1996), Levelt/Roelofs/Meyer (1999), and many others. The importance of the syllable has also been recognized in language acquisition, in particular in child language acquisition. At an early age, children recognize the syllable as the basic perception unit. The syllable is also relevant in aphasia research, where it has been shown that there seem to be a speaker's sensitivity for syllable patterns and sonority, the both of which are lost only very late in the course of the illness (cf. Berg 1992, Stenneken et al. 2005).

One question discussed again and again in linguistics regards the general relevance of the syllable as a linguistic unit and its position within theoretical linguistics. Since it is undoubted that phonology cannot be done without the syllable as the basic articulatory and perception unit, the cognitive status of the syllable is in the

focus of ongoing discussions. The main question is to which extent the syllable plays a role on the semantic level and in language processing. There is psycholinguistic evidence for the cognitive relevance and for an internalized knowledge of the syllable structure by L1-speakers, this being of relevance for any language production and reception model. Even though a semantic and cognitive status (syllables hardly ever carry lexical meaning) can be disputed, it remains quite clear that linguistics cannot dispense with the syllable, since it is the most important frame of phonological and phonetic processes, and the basic unit and constituent of any hierarchically higher unit (morphemes, words, lexemes).

1.3 The syllable as a linguistic unit

Linguistics is usually not the proper place for a discussion of ontological issues. Having in mind related references on the syllable, one could at least partly get the impression that in some cases, a lot of effort is put into the question about the "reality" of the syllable as an ontogenetic category. However, we believe that searching for the "real" essence of a linguistic unit is, in a strict sense, unproductive, and even unnecessary. An adequate alternative to an ontogenetic approach is to focus on the question of a proper definition, based on terminological conventions and detailed criteria.

The complexity of a syllable definition is obviously biased by the fact that it is one of the few linguistic units or categories which are more or less intuitively perceivable by a native speaker of a language (what easily can be proved by the ability of chanting and declaiming, and by the intuitive recognition of rhymed patterns in poetry). However, the intuition does not help to identify this unit unanimously and gives no information about setting the borders of this unit (= syllabification).

The identification and definition of the syllable is the core task of linguistics, and the overall relevance of the syllable results from a set of criteria, summarized by Altmann (1996, and based on Salthe 1995). According to them, a linguistic entity can be considered a linguistic unit if it (1) can be (operationally) isolated from its environment relatively well. The isolation implies the identification of boundaries, which is related to the used grammar, the context, the research question analyzed, etc. However, in many cases a bit of vagueness, ambiguity, and fuzziness can remain, even when setting up dozens of criteria. (2) Therefore, one minimum requirement is that a linguistic unit has an identity – at least a vague one. A simple empirical proof of it is to have a look at the historic development of a linguistic unit. A unit can either remain steady or it changes, but in any case, it should not disappear. (3) A linguistic unit should take part in at least one (synergetic) control cycle. To put it into more general terms, the unit is not an isolated one, but it interacts with other units and/or it can influence other ones. Moreover, a proposed unit should (4) meet the requirements of the members of a language community. Taking into consideration this catalogue of criteria, one has to emphasize, in particular, the importance of the syllable as a unit in natural language processing – both for the speaker during encoding, and for the hearer during decoding linguistic information. As already pointed out above, there is a plenty of evidence for the "cognitive" relevance of the syllable in language processing.

Based on these considerations, it remains quite clear that the question of a proper syllable definition is indeed not an ontological one, but rather a methodological

and theoretical one. In quantitative and synergetic linguistics, a focus is laid on the question to what extent the syllable participates in shaping the overall structure of the linguistic system and in interrelating with other units.

1.4 Principles of segmentation

The definition of the syllable is in many ways related to the theoretical framework one relies to. In order to give an overall idea about syllable definitions discussed in the past, a brief overview on important attempts and conceptions is presented in the following section.

One basic attempt is to take into consideration the physical substance or "material" characteristics, helping to identify the syllable. Among others, the sonority (i.e., the amplitude) of segments, the opening and closing of the mouth cave, the breathing stream, and more generally muscle impulses (cf. Stetson 1951, Kelso/Munhall 1988) have been discussed as being relevant for its identification.

Regarding more sophisticated linguistic criteria, there are at least two main competitive approaches, which help to identify the syllable and the syllable borders. One is based on phonotactic considerations, popular, in particular, in the realm of structuralism(s), and the other one is based on the principle of sonority. The latter is relevant in natural phonology, optimality theory, and many other approaches, which are influenced by a more processual way of linguistic thinking.

To give a brief insight into the phonotactic approach, one can rely on Pulgram (1970), one of the most influential monographs on the syllable and syllable division from a structuralist point of view. His basic idea is that the syllable is shaped by the same patterns as the word-initial and word-final occurrences of phonemes and phoneme combinations are:

> "[…] the first syllable of a cursus, nexus, or word has the same phonotactic constraints at its beginning as does the word. By the same token the equation prepausal = cursus-final = nexus-final = wordfinal can be extended by adding: syllable-final. This establishes that the last syllable of cursus, nexus, or word has the same phonotactic constraints at its ends as does the word" (Pulgram 1970: 45).

Based on this criteria, a tentative segmentation of syllables can be performed. However, some additional principles are required for a better and clear segmentation, among others: (1) the principle of the maximal open syllabicity, which results in a preference for open syllables (ending with vowels); (2) the principle of the minimal coda and maximal onset (i.e., onset structure is preferred); and (3) the principle of the irregular coda, where it is stated that any occurring irregularity is more likely to occur in the coda than in the onset.

One disadvantage of this approach is its close relatedness to the concept of word and its focus on the word-initial and word-final structure, being in particular problematic for languages without word-similar units. However, one major advantage of Pulgram's approach is its principal openness towards empirical applicability. What is particularly interesting is a further modification by Lehfeldt (1971: 221), who suggests to implement the frequencies of word-initial and word-final combinations –

which allow to distinguish between marginal and non-marginal phoneme clusters (cf., for an application on Russian syllable segmentation, Kempgen 1995) – into the syllabification process.

The second most important feature for syllable segmentation is the concept of sonority, also called sonority sequencing principle or consonantal strength. The basic idea goes back, among others, to Sievers (1885) and Jespersen (1904), who distinguished various subclasses of vowels and consonants according to their degrees of sonority. Furthermore, it has been observed that in syllables, the sonority rises the nearer it comes to the nucleus, followed by a gradual decrease after the nucleus. In the past, many different sonority scales have been proposed (Vennemann 1972, Foley 1972: 97, Ladefoged 1975, Hooper 1976, and many others), which differentiate from each other in some minor aspects only. Roughly, a hierarchy of sonority begins with vowels, which are followed by liquids, nasals, fricatives, and stops, these having the lowest degree of sonority. Thus, sonority seems to determine the internal positioning of segments within the syllable, belonging to various phonetic subclasses. However, it has been noted that sonority does not help in determining syllable borders in all cases clearly, since sonority plateaus or irregular positionings of segments can also be observed. To conclude, it appears that sonority is a general principle, responsible for the segmental shape of the syllable, but it cannot be operationalized in such a way that an exact segmentation is achieved.

1.5. Quantitative analysis of the syllable: A synergetic approach

The syllable as a linguistic unit plays a crucial role in quantitative linguistics. First of all, the syllable is understood as the direct constituent of the word. This makes understandable why the syllable is used quite often as a measuring unit in word length studies. Moreover, the quantitative properties of the syllable are of interest on their own. The syllable also plays a particularly prominent role in Menzerath's Law, where, among others, it is stated the longer the word is, the shorter the syllable is, or the longer the syllable is, the shorter the sound duration is (see Cramer 2005 for further details).

It has been outlined above that the principal relevance of the syllable is proven, as it can be part of a network of mutual interrelations with other linguistic properties and units. Recently, there is no full-fledged synergetic control cycle for the syllable available, but mostly some tentative ideas and fragments. A first attempt goes back to Zörnig/Altmann (1993), where it is asked by which properties the number of canonical syllable types (syllables are noted as sequences of vowels (V) and consonants [C]) is determined. They focus on four selected properties:

1. The phoneme (or grapheme) inventory, which is at one's disposal and participates in the construction of syllables.
2. The vocabulary size of one language, which is required within one language for fulfilling communication needs.
3. Restrictions regarding the phoneme distribution, since it is well-known that not all possible phoneme combinations are realized, but only a small subset.
4. The syllable length, which – as it is well-known from Menzerath's Law – stochastically depends on the word length.

Fig. 1.4 gives a graphical representation of the stated interrelations, where the mutual dependencies between the variables can be seen. The control cycle includes the most important factors, influencing the number of syllable types in a language.

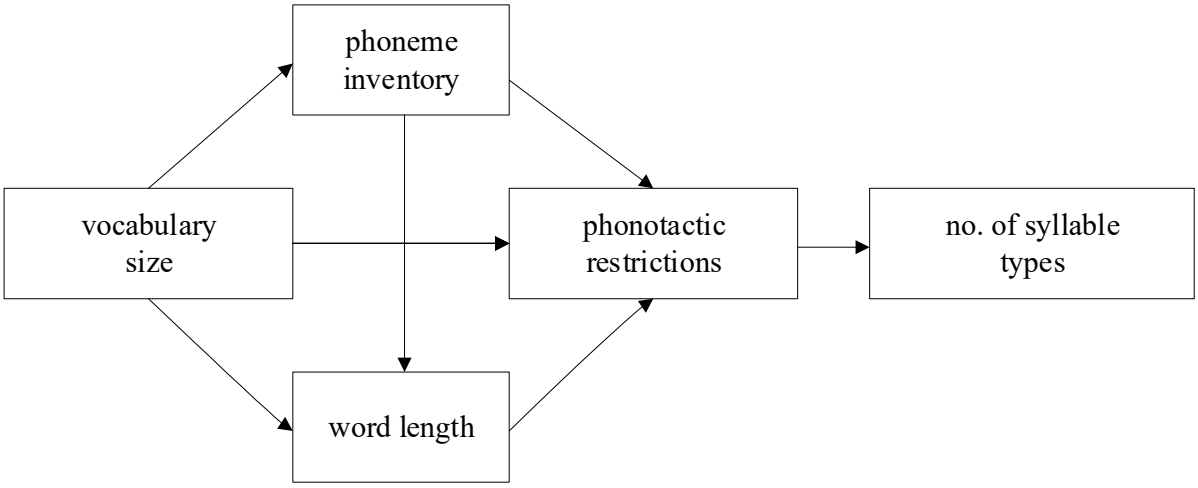

Fig. 1.4. Synergetic control cycle: no. of syllable types (Zörnig, Altmann 1993)

A second attempt to develop a synergetic control cycle with the syllable at its core goes back to Kelih (2012). His basic idea is to leave aside general properties like the phoneme inventory or the vocabulary size (both characteristics are indeed multiply correlated and interrelated with word or syllable lengths), and to focus much more on the syllable level and characteristics and properties closely related to it.

It is again Menzerath's Law (the longer the word, the shorter its syllables) that appears to be the most important factor shaping the syllable structure. This basic law has an overall impact on many other syllable-related properties, which is also reflected in the proposed schema of interrelations, which are discussed in detail below.

(1) Since the syllable length depends stochastically on word length, it can be derived deductively that the overall syllable structure and the syllables types in 1-, 2-, 3-, 4-, …, x-syllable words (henceforth, word length classes) have a level-particular shape, too.
(2) In syllable studies, the canonical syllable type – i.e., the notation of a syllable as sequence of consonants (C) and vowels (V) – is an important heuristic tool. Moreover, based on this notation, the overall complexity of syllables can be caught easily. On the basis of Menzerath's Law, one can state an interrelation between the number of canonical syllable types and the word length class – there are less canonical syllable types in higher word length classes, since the syllables are shorter in these words. Thus in one-syllable words, a high number of syllable types should be observed.
(3) In addition to the number of syllable types, the frequencies of canonical syllable types have to be taken into consideration as well. Since the frequency plays an outstanding role in almost every synergetic approach, regarding the syllable, at least two kinds of frequency have to be distinguished: (a) frequency of individual syllables and (b) frequency of canonical syllable types. The latter is in the focus of the presented study, where mostly the question of modelling is tackled (see section 1.7). Coming

back to the frequency, at least two hypotheses have to be mentioned: (c) the longer (= more complex) the syllable, the lower its frequency; and (d) the longer (= more complex) a syllable type, the lower its frequency. Both relations should by modelled by some kind of a power law. However, it remains unclear whether the length is a function of the frequency, or the other way round. This has to be determined empirically.

(4) Restrictions on phoneme distribution are a further important influence factor shaping the syllable structure. For the sake of simplicity, at the level of phonotactic restrictions only the number of consonant combinations in the onset and coda are taken into consideration in Kelih (2012). Following the "Onset-First-Principle", more combinations should be found in the onset than in the coda. In any case, this hypothesis depends on the chosen syllabification procedure and language-specific peculiarities.

(5) There should be an interrelation between the number of consonant combinations and the number of syllable types in different word length classes. Consequently, this should also be the case for the word lengths. The more restriction is at work in the onset and/or in the coda, the fewer syllable types can be processed.

(6) In syllable phonology for the coda and onset, a so called mirror effect (Vestergaard 1967, Sigurd 1955) has been observed. As a tendency, the consonant combinations found in the coda are the reverse (mirrored) forms of the onset – i.e., "C_1C_2" in the onset appears in the coda as "C_2C_1". This phenomenon can be explained by the sonority hierarchy principle, which is responsible for the internal pre- and post-nucleus positioning of consonants. The proposed mirror effect increases the symmetry in the syllable system, which, in return, has an influence on the number of different syllable types. In the book, this tendency will be investigated in Chapter 5.

(7) The aforementioned properties are mainly related to the segmental level. However, the syllable also plays an important role in the prosody and intonation, and that is why at least some properties of this level should be integrated in the research, too. Available supra-segmental features have an impact on the word length, since they can help to reduce lengthening processes. Moreover, it has to be considered that the question of a proper quantification seems to be at the beginning and that the type of accent (pitch accent, stress, tone, etc.) is directly related with the syllable structure of languages. As one possible empirical treatment of these problems, the unaccented and accented syllables can be taken into account, which opens the door to an overall analysis of the rhythmic organisations of language systems.

The proposed framework (cf. Fig. 1.5) is to be understood as a first tentative attempt to a fully fledged synergetic syllable theory and should be extended with other syllable properties, characteristics, and features of the phonological and morphological levels.

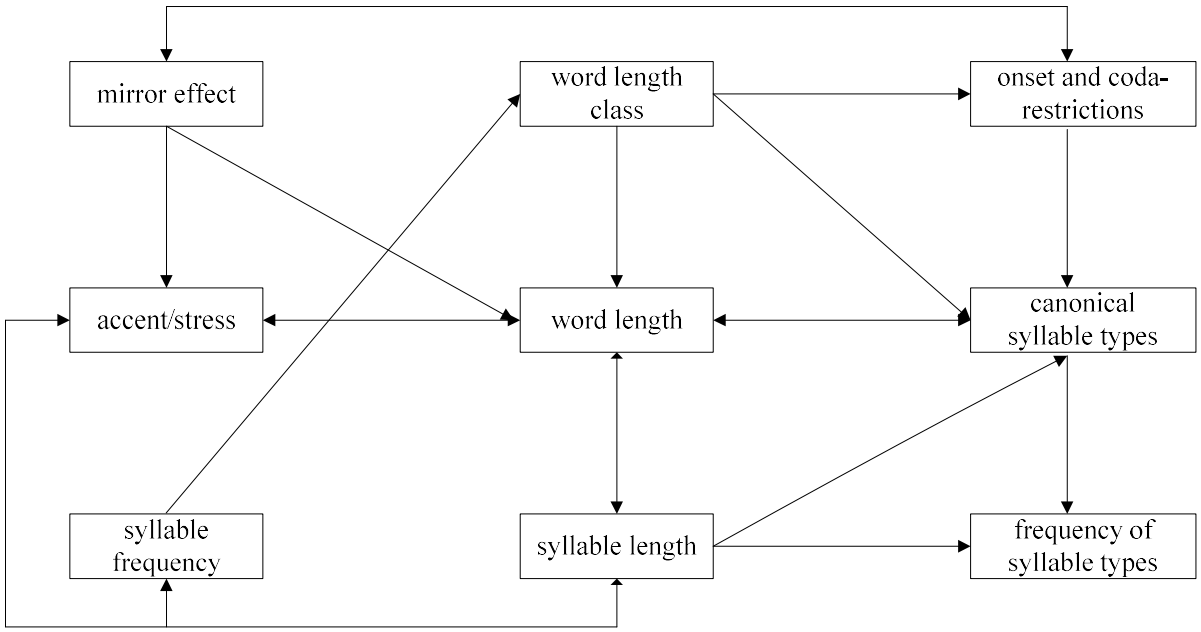

Fig. 1.5. Extended synergetic syllable control cycle (according to Kelih 2012)

In the future, the proposed tentative control cycle can be modified and specified; in particular, it is well-known that the syllable can even be related with the overall grammatical and syntactic structures of languages (cf. Fenk-Oczlon, Fenk 2005, 2008).

1.6 Generalities on quantitative research

Usually, we devise new concepts for entities in order to say "what is there" – e.g., in a text, there are sentences, clauses, phrases, words, morphemes, syllables, phonemes, parts of speech, etc. In physics, one defines new concepts, expresses a property of them using mathematics, but the empirical finding of a real counterpart may take years. In linguistics, a "wrong" definition of concepts may lead to a new direction, but if after a long time no background theory is found – i.e., no hypotheses are positively tested -, the new discipline dies. One tries to save it by redefinitions and new data, but without a possibility of testing, the problems of "how it is" and "how it behaves", the unit falls into oblivion.

The first question is usually answered by proposing a quantification containing other concepts and inserting all into a measurement prescription. One tries to measure the phenomenon, but without answering the second question, it is not possible to perform further steps. Usually, the second question requires a background theory from which the behaviour of the phenomenon is derived. One sets up hypotheses and tests them, using data from one language at first, then from other ones. A hypothesis may hold true only if it can be tested in all languages – but this is never the case and, as a matter of fact, it is impossible. Further, no linguistic phenomenon is isolated, there are always some other phenomena connected with it and influencing its behavior. That means that there never exists a completed theory in linguistics (as a matter of fact, in no science). The greatest linguistic step has been done by Köhler (1986, 2005), who presented the synergetic self-regulating circle, which waits for its extension. This way of thinking is based on Zipf's previous investigations (1952). The history shows that

language phenomena had been analyzed quantitatively already earlier (cf. Köhler 1995), but, without supporting the research with a theory.

If our hypotheses concern the simple form of a phenomenon, we have to do at least with its (immediate) components; and the components may also be parts of other systems. They themselves have, possibly, their own components, can be classified in many ways, etc. The number of ways is infinite. The same holds for supra-systems. First, phenomena can be ascribed to some classes, the new classes belong to super-classes, etc. Language, just as the rest of the nature, is not described by its "highest" or "lowest" level because these are unknown. The subdivision into "langue" and "parole", or into "competence" and "performance", "synchrony" and "diachrony" are merely the first trials to find an orientation. The majority of classes and levels are (for us) probabilistic; for example, every pronunciation of a sound is different from the previous one or from the pronunciation by another person. In modern science, we speak rather about systems and use systems theory for solving a problem. Now, the mathematical models we derive from a background theory do not represent the "truth", but enable us to use the result for further derivation, hypotheses construction, and necessary testing. Every linguistic hypothesis is corroborated only to a certain degree. However, if we accept it, it must hold true for all languages. The differences among languages should be contained in different parameters of the models, perhaps different connections with other properties, but the respective functions may originate in various differential or difference equations.

We shall always find "exceptions", e.g., one of the classes deviates strongly from the trend presented by the other classes. In that case, the modelling may be adapted, for example by adding a separate class given by separate parameters – e.g., $y_1 = \alpha$, and the others as $y = f(x)$. If one models probabilistically, one must care for the correct sum of probabilities, yielding 1. In all languages, there are some "exceptions" caused, e.g., by borrowings, but one can omit them, if necessary and possible. Moreover, the evolution of a language creates exceptions, too – for example, if a class changes and loses its members, the remaining ones must be considered exceptions. This is the case, e.g., with strong and weak verbs in German or English: the class of strong verbs changes, it loses its members, which pass to the weak class, and the rest will be, in the future, considered exceptions. It needs sometimes centuries until a change is complete.

There are always several possible models for the same phenomenon. One can perform a choice adhering to the following principles: (1) One may set up a probabilistic or a functional model, or one may choose a continuous or a discrete model. This is possible because reality is neither continuous nor discrete, and the functional or probabilistic dependencies are merely our views, our trials to "make order". (2) One should use the simplest function expressing adequately the data – i.e., a function with as few parameters as possible. The parameters are some (necessarily) interpretable properties or requirements, or forces whose interpretation makes the model acceptable and useful for further research. In linguistics, one frequently uses the Zipfian-Köhlerian requirements, e.g., easy pronunciation, easy comprehension, easy storing (cf. Köhler 2005). (3) If possible, one should avoid polynomials, due to various reasons: (a) they have usually too many uninterpretable parameters; (b) they are not easy to be subsumed under a theoretical roof; (c) they are able to capture any sequence, but do not always yield an explanation; (d) sometimes they have more

parameters than there are classes, etc. (4) One should avoid the "normal" (Gaussian) distribution because nothing in language seems distributed normally; there are a number of requirements that support asymmetry. Nevertheless, everything can be "normalized" by a correct transformation.

Comparisons of text types, languages, authors, texts, etc., can always be performed using a statistical test. Here, one can use either the complete numerical series, or its indicators such as moments, or one can rank the data and perform ranking tests. The same can be done for dictionaries, and also when one studies the changes from, e.g., Latin to French, the differences between cognate languages, etc. With testing, we currently apply some usual tests based on normality, which is nothing "criminal" (though nothing is distributed normally in language) because the test cares for previous "normalization". Statistical tests are our first steps towards the confirmation of a hypothesis.

Sometimes, the question "what is the phenomenon?" cannot be answered directly because for us it is a concept ascribed to some data. The definition should merely help us to identify its existence in texts or dictionaries. If we speak about syllables, we can find a definition which is not equal for all languages. In some languages, one uses, e.g., the term "mora". One has problems with stating the boundaries of the element, and even the counting results obtained by computers must be corrected sometimes. In many languages, one has problems with diphthongs, in other ones with syllabic consonants, sequences of consonants, foreign syllables, nasal vowels, weak vowels, etc. Many definitions are merely conventions introduced by linguistic schools. Reading the literature about syllables in individual languages, one always finds different segmentation rules; hence, even native speakers have problems. The prescriptions for the hyphenation of words hold rather for the written language than for the spoken one, but in no case do they hold for syllable division in non-alphabetic languages. While in agglutinative languages the syllable boundaries mostly coincide with morphological boundaries, in inflectional languages it needs not be so.

However, in any case, the general line can be followed. In the present book, we shall analyze the syllables in some Slavic languages using the same (translated) text of the first chapter of the Russian book *Kak zakaljalas stal'* ("How the Steel Was Tempered") by Nikolai Alekseevich Ostrovsky. The same comparison will be performed with the translations of the Hungarian poem *Szeptember végén* ("At the End of September") by S. Petöfi, and a number of texts taken from various languages should help us to find some common regularities. The other texts represent the situation in the given language, for the given author, and for the given individual piece of writing.

1.7 Modelling

Syllables have been analyzed frequently, and both their types and their lengths are no new problems. One tried to approach the problem using probability distributions, e.g., the Conwell-Maxwell-Poisson distribution; here, we shall apply simple functions and show their adequacy in several languages. The syllable types, when ranked according to their frequency, abide by the Zipf-Alekseev function, defined as

$$y = cx^{a+b\ln} \quad ,$$

usually with added 1, which is sometimes necessary because the frequencies cannot be smaller than 1. In the differential equation, it simply means that the change of y depends on the previous value, $y-1$ – i.e., we consider the relative rate of change as

$$\frac{y'}{y-1}.$$

Needless to say, there are many other functions expressing this regularity quite well; we shall try to find a unique one. Nevertheless, many times (in some individual languages), the exponential function is sufficient for capturing the trend. It has the advantage of containing merely two parameters.

The length of syllables given in the number of phonemes abides either by the Lorentzian function (cf., e.g., Andreev, Místecký, Altmann 2018) defined as –

$$y = \frac{a}{1+\left(\frac{x-b}{c}\right)^2},$$

or by the Menzerathian function defined as –

$$y = ax^b e^{-cx},$$

both of which can take a parabolic form. All of them have been many times derived in the linguistic literature. The substantiation of the Menzerathian function is linguistically much easier than that of the Lorentzian, and the fact that the word length and the lengths of other linguistic entities abide by it, too, is a further reason for testing and – in the positive case – accepting it. The Menzerath law holds true for the immediate components of higher units (cf. Altmann, Schwibbe 1989), but here, we shall show that it holds very generally, at least for the length of syllables.

In every language, some problems arise, but in any case, the analysis of a text follows some prescriptions written for the given language by linguists, and one does not make an error if one follows them.

As to the comparison of languages or texts, one may apply, e.g., the chi-square test for frequencies, or a non-parametric rank-test for ranks. Here, a plethora of problems seems to be opened. Each aspect (types, lengths, asymmetry, open/closed syllables, relations to grammar, distances between equal entities, etc.) can be compared, and evaluated, especially if it is expressed formally.

Although syllables are no grammatical or semantic phenomena, their study can be theoretical, too. One tries to find regularities, which may be restricted to a given language or language family and, at last, one inserts all respective phenomena into a theoretical framework. Hence, our aim is not only classification or typology, but a search for regularities, which can obtain the status of laws later on.