History of Quantitative Linguistics

Since a historiography of quantitative linguistics does not exist as yet, we shall present in this column short statements on researchers, ideas and findings of the past – usually forgotten – in order to establish a tradition and to complete our knowledge of history. Contributions are welcome and should be sent to Peter Grzybek, peter.grzybek@uni-graz.at.

Historical Remarks on the Consonant-Vowel Proportion – From Cryptoanalysis to Linguistic Typology

The Concept of Phonological Stoichiometry (Francis Lieber, 1800-1872)

Peter Grzybek

Studies on the frequency of vowels and consonants in general, and on consonant-vowel proportions specifically, have a history which reaches back much longer than is usually assumed. Unfortunately, we know this history only most rudimentarily.

The beginning of such studies can be seen, it seems, in early cryptographical and cryptological literature (cf. Ycart 2013). Arab philosopher and mathematician Yaʿkūb ibn Ishāq Al Kindī (800-873), for example, in his manuscript *On Deciphering Cryptographic Messages*, which seems to be the oldest known description of cryptoanalysis by frequency analysis, clearly points out the different frequencies of vowels and consonants, and arrives at the conclusion that "the number of vowels in any language would be greater than non-vowels".[1]

Six centuries later, in the context of West European Renaissance, and ignorant of his Arab predecessor, Italian humanist Leon Battista Alberti (1404-1472) started his relevant ruminations in his *De Componendis Cifris* (ca. 1466-67); he arrived at the conclusion that if we take "one or two pages of poetry or prose and extract the vowels and consonants, listing them in separate series, vowels on one side and consonants on the other, you will no doubt find that there are numerous vowels"[2]. Moreover, what is even more important in our context, is Alberti's attempt to quantify the CV relation:

> From my calculations, it turns out that in the case of poetry, the number of consonants exceeds the number of vowels by no more than an octave[3] [non amplius quam ex octava], while in the case of prose the consonants do not usually exceed the vowels by a ratio greater than a sesquialtera [ferme ex proportione quam sesquitertiam]. If in fact we add up all the vowels on a page, let's say there are three hundred, the overall sum of the consonants will be four hundred.[4]

[1] Quoted after Ycart (2013: 1)
[2] Quoted after Ycart (2013: 1)
[3] There has been a debate as to the meaning of "an octave", but it seems reasonable to side with Ycart's (1012) argument and interpret it in terms of "one eighth".
[4] Quoted after Ycart (2013: 9)

The distinction of vowels and consonants on the basis of frequency became more or less common during the following centuries and, in fact, one of the standard procedures in crypto-analysis. There is no need to go into historiographic details here, concentrating rather on the vowel-consonant proportion specifically, which has been discussed to a much lesser degree. In this respect, Alberti was much more concrete as to numeric details than most of his followers, who nevertheless made concrete suggestions as to the CV proportion. One of them was David Arnold Conrad who, in his 1732[5] *Cryptographia denudata, sive Ars Decifrandi*, gave such an estimation of the CV relation: "The Vowels, generally five, are four times out-numbered by the Consonants, the Vowels must therefore recur most frequently" [Quoted after Ycart (2013: 6)]. Yet another estimation of this kind, containing a quantifying assertion about CV frequencies, can be found in Christian Breitenhaupt's 1737 *Ars Decifratoria*:

> The frequency of letters should be noted in general, since in any language vowels are more numerous than consonants. The reason for making these observations is obvious. Actually, for a given number of vowels, the corresponding number of consonants must be larger by five fourths; it cannot be otherwise, vowels being more frequent than consonants.[6]

Here is not the place to go into more details as to the history of studies on CV proportions. It seems that, from a historical point of view, and going beyond the narrow field of crypto-graphy, studies in this field have been epistemologically motivated by three major realms of interest:

1. *Genuinely linguistic*. The primary field of interest is, of course, linguistic: after all, it is a genuinely linguistic issue to define consonants and vowels, as well as other units; subsequent questions have mainly concentrated on typological issues, with regard to intra-lingual aspects (which factors have, within a given language, impact on the CV proportion?) as well as to inter-lingual and cross-linguistic aspects (is the CV proportion a possible characteristic for language typology?)
2. *Aesthetic and poetic*. In this respect, a leading question has been, is it possible to define phenomena like the euphony, or harmony, of a given language, or of individual texts in that language, on the basis of the CV relation?
3. *Pedagogic*. Here a major issue has been the question, if knowledge about the CV proportion can help in defining matters of text difficulty and understandability, or distinguish "easy-to-learn languages" from more (or less) "difficult" ones, and related questions.

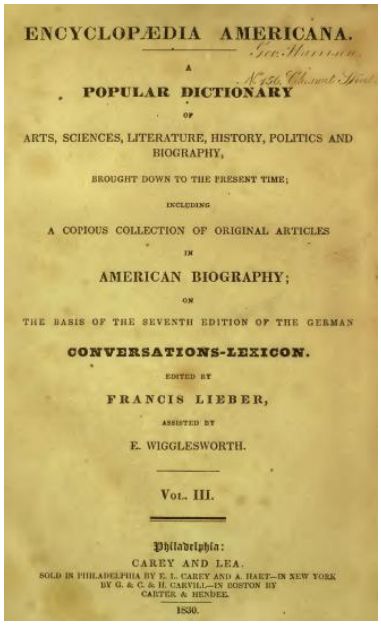
As a matter of fact, any answer to the second and to the third question must, in one way or another, start from an at least implicit assumption concerning the definition of the basic terms; in case the definitions are explicit, they depend, historically speaking, on the state of linguistic knowledge and, from a contemporary point of view, on the concrete linguistic theory chosen. In any case, the underlying definition is subsequently relevant, of course, for the frequencies of the distinguished items, as the basis for calculating the proportion between them. Quite naturally, genuinely linguistic approaches to the CV issue increased and have been prevailing in the in the 20th century, with the rise of linguistic theory. In this respect, mainly structuralist and typological approaches, as e.g., Isačenko's (1939/40), Krámský's (1946/48), or Skalička's (1966) seminal papers; these approaches have later been thoroughly reflected by Altmann and

[5] Conradus' texts were re-published 10 years later, in 1742, in the *Gentleman's Magazine*, in a series of articles.
[6] Quoted after Ycart (2013: 1)

Lehfeldt (1973) from a methodological point of view. And despite some justified critique, as brought forth e.g., by Kempgen (1991), one can only agree with Kelih (2010) stating that the basic idea about CV proportions in languages still today is of great relevance and timeliness.

Nevertheless, the historical interest in CV proportion is much older than these approaches, and it is just this historical background about which our knowledge is but fragmentary, from a historiographical point of view. Scholars, who were interested in related issues at the beginning of the 19th century, more often than not came from disciplines other than linguistics, the latter more or less in the *status nascendi*, rather than being an established branch of science. Among those scholars was, as has been shown elsewhere (cf. Grzybek 2006), Czech zoologist and mineralogist Svatopluk Presl, who not only presented one of the earliest Slavic letter statistics, but also considered to CV proportion to be an index at a language's euphony and degree of learning difficulty.

Among those early scholars interested in CV proportions also was the German-American Francis Lieber, whose linguistic contributions have almost been forgotten by today's linguistic audience.


ENCYCLOPÆDIA AMERICANA.

A

POPULAR DICTIONARY

OF

ARTS, SCIENCES, LITERATURE, HISTORY, POLITICS AND BIOGRAPHY,

BROUGHT DOWN TO THE PRESENT TIME;

INCLUDING

A COPIOUS COLLECTION OF ORIGINAL ARTICLES

IN

AMERICAN BIOGRAPHY;

ON

THE BASIS OF THE SEVENTH EDITION OF THE GERMAN

CONVERSATIONS-LEXICON.

EDITED BY

FRANCIS LIEBER,

ASSISTED BY

E. WIGGLESWORTH.

Vol. III.

Philadelphia:

CAREY AND LEA.

SOLD IN PHILADELPHIA BY E. L. CAREY AND A. HART.—IN NEW YORK BY G. & C. & H. CARVILL.—IN BOSTON BY CARTER & HENDEE.

1830.

Francis Lieber (March 18, 1800 – October 2, 1872), originally known by his German name Franz Lieber, was a German-American publicist, jurist, and political philosopher. Born in Berlin, he joined the Prussian Army during the Napoleonic Wars, and was wounded at Waterloo. Returning to Berlin after the Napoleonic wars, he attempted to pass the entrance exams to the University of Berlin; but being member of the Berliner Burschenschaft, a student fraternity, inspired by liberal and nationalistic ideas, which opposed the Prussian monarchy, he was denied admission. Moving to Jena he matriculated in 1820 to the University of Jena, and within a span of four months finished writing a dissertation in the field of mathematics. As the authorities caught up with him, he left Jena for Dresden to study topography, but as soon as the Greek Revolution of 1821 broke out, he volunteered his services. Lieber left Germany forever in 1825; for a short time he resided as a teacher in London, and in 1827 he embarked for the United States. During the next five years, during his residence in Boston (1827-32), he was occupied with the compilation of the 13-volume *Encyclopedia Americana: A popular dictionary of arts, sciences, literature, history, politics and biography, brought down to the present time; including a copious collection of original articles in American biography*. The encyclopedia was based on the 7th edition of the German *Brockhaus Conversations-Lexicon*, which had appeared in 1827 under the title *Allgemeine deutsche Real-*

Encyklopädie für die gebildeten Stände (Conversationslexikon), and which was, after *Dobson's Encyclopædia* (1789–1798), the first significant American encyclopedia.

In 1835, Lieber moved to Columbia, S.C., where he occupied the position of professor of political economy in the South Carolina College for twenty years; and here he produced his most important works: *A Manual of Political Ethics* (1838); *Legal and Political Hermeneutics* (1839); and *Civil Liberty and Self-Government* (1852). In 1856, he was called to Columbia College, New York, to take the chair of political economy, and in 1860 accepted the chair of political science in the Columbia Law School, giving up his chair of economics. He was the author of the *Lieber Code* during the American Civil War, also known as *Code for the Government of Armies in the Field* (1863), which laid the foundation for conventions governing the conduct of troops during wartime. Lieber died in New York, September 2, 1872.

Among Lieber's numerous works are a number on language, only three of them having been published, however. The most important in this respect are considered to be his 1837 article "On the Study of Foreign Languages", his 1850 contribution "On the Vocal Sounds of Laura Bridgeman", and his 1852 "Plan of Thought of the American Languages".

In the first of these articles, Lieber defended the teaching of the classical languages at schools; it is Lieber's discussion of the nature of Native American languages that had a lasting influence. In this article, as well as in the shorter one from 1852, Lieber praised native American languages, compared them favorably with the classical languages, and coined the term 'holophrastic' to describe their agglutinating or polysynthetic nature. Lieber explained this phenomenon in his 1837 and 1852 articles, discussed previous terms used to describe it (including agglutinative and polysynthetic), and explained why he considered his coinage 'holophrastic' to be a superior term, which indeed was used in works on Native American languages during the remainder of the nineteenth century.

Lieber's 1850 article is about the vocal sounds of Laura Bridgman, who is known as the first deaf-blind American child to successfully gain a significant education in the English language, some fifty years before the more famous Helen Keller. Bridgman was taught tactile finger spelling becoming completely fluent in it; she was not taught oral language and could only make a limited range of vocal sounds. She could communicate rapidly with anyone else who knew tactile finger spelling. Lieber's article was unique for its time.

In addition to the three published articles described above, and in addition to various unpublished articles, most probably many language-related articles in the *Encyclopedia Americana* were written by Lieber, although the contributions to the *Encyclopedia* were unsigned. One of these contributions to the third volume (1830) is on "Consonants". Among others, a number of calculations concerning the CV proportions of different languages are presented for comparative purposes, and it seems that these statistics, along with their interpretations, are among the earliest of this kind we know of.

According to Lieber (ibd., 450), the "various interesting relations of consonants to vowels, and of the sounds and letters in the different idioms, have not yet received any satisfactory investigation […]."

For Lieber, the study of euphony, or harmony, was a central concern in his cross-linguistic analyses of CV proportions: far away from saying that the euphony of a language depends entirely on this proportion (ibd. 450), Lieber was convinced of the fact that the "melodious sound or music of a language depends, in part, upon the proportion of the vowels to the consonants, a language becoming too hard if there are too many consonants" (ibd.).

In order to establish the CV relations of different languages, Lieber did not base his analyses on the paradigmatic level of inventories, but took passages from different texts, thus integrating the analysis of frequency of occurrence. Again, the author was well aware of a number of possible methodological problems. Attention was paid, among others, to text size: "The different passages were very similar in size, so that the number of syllables in each

would be very nearly the same." And although the author could not pay attention to further possibly intervening factors, he was at least well aware of the fact that the choice of different text types might result in possible differences: "To give anything like accuracy to such investigations, it is obvious that the results ought to be taken both from prose and poetry, also from many different writers, and the language of conversation" (ibd., 451).

For English, Italian, German, Portuguese and Spanish texts, three stanzas were taken from each of the following poems: the beginning of Lord Byron's *Childe Harold*, Torquato Tasso's *Gerusalemme Liberata*, Goethe's *Zueignung* (prefixed to his *Faust*) the *Luisiada* by Luís Vaz de Camões, and the Spanish epic poem *La Araucana by Alonso de Ercilla.* For French, he took 24 lines of the beginning of Racine's *Thébaïde*; for Greek (Ionic), 24 hexameters of the beginning of the *Odyssey*, and for the Attic dialect, the beginning of the *Anabasis*; for Latin, the 24 first hexameters of Ovid. In addition to these languages, Lieber offered data for Hawaiian (still termed Sandwich islands in the 18th century tradition of James Cook), Seneca Indian, Chahta Indian, Sanscrit, Malay, Persian, Hebrew, and common Arabic.

For some languages, Lieber reported separate results for what he termed 'orthographic proportion' vs. 'phonic proportion': those languages which he assumed to be characterized by an approximately 1:1 sound-letter relation, are counted by letters, all others by sounds (a sound possibly being represented by more than one letter).

Starting with an analysis of the *Odyssee*, a text in the Ionic dialect of Greek, Lieber found a CV proportion of 3:4, which he considered to be "a very melodious proportion" (ibd. 451). Comparing it to the results for the Attic dialect, for which he found a CV proportion of 1:1.006, he stated a difference of 0.327:

Ionic	=	3 : 4	=	1 : 1.333
Attic	=		=	1 : 1.006

				0.327

Similarly comparing Latin (with a CV relation of 6:5) to Italian (11:10), he found a difference of 10% between both languages. Table 1 contains the results for all languages as represented by Lieber, the data marked by an '*' being based on what Lieber termed 'phonetic proportions'; additionally, the last column contains the CV quotient based on the data given.

By way of a conclusion, Lieber arrived at the result that not only languages seem to be characterized by different CV proportions, but also do languages belonging to a common family seem to follow similar patterns. According to the author, it can easily be seen "that, in the languages of Latin origin, the proportion of consonants to vowels is much smaller than in the Teutonic idioms" (ibd., 452). But he was well aware of the pioneering state and limited reliability of his analyses and results. With due caution he frankly admitted that "the conclusions […] are rather to be regarded as indications of what might be learned from more thorough inquiries, than as facts from which general deductions can be safely drawn" (ibd., 451).

As a consequence, instead of jumping to hasty conclusions, Lieber (ibd., 452f.) suggested some kind of research program, including tasks as the following:

> "to compare the proportions of consonants to vowels, in such different families of languages; to show the proportions of the gutturals, labials, &c., of the different idioms; and, again, the proportion of these letters in the various families of languages, or according to the different parts of the earth to which they belong, as Asiatic, European, &c. languages, and many other calculations."

Table 1

Consonant-Vowel proportions for different languages (Lieber 1830)

		C	**V**
Sandwich islands		1	1.800
Greek	Ionic	1	1.333
	Attic	1	1.006
Portuguese		1.020	1
Common Arabic		1.080	1
Italian		1.100	1
Seneca Indians		1.180	1
Chata Indians		1.200	1
Sanscrit	*	1.200	1
Latin		1.200	1
Hebrew	*	1.200	1
Spanish		1.240	1
Persian	*	1.330	1
Malay	*	1.330	1
French	*	1.340	1
	orthographic	1.270	1
Dutch		1.500	1
English	*	1.510	1
	orthographic	1.520	1
Swedish		1.640	1
German	*	1.700	1
	orthographic	1.640	1

In his concluding remarks, But Lieber (ibd., 453), methodologically generalized and embedded his approach, referring to Duponceau's (1818) ruminations on English phonology:

> "We have no doubt that the more the science of languages is developed, the more obvious will be the necessity of the study of *phonology* […] the knowledge of the sounds produced by the human voice." And he was, on the one hand, a child of his time, but much ahead of his time, on the other, when he compared the contours of this field of phonology to be developed to contemporary approaches in chemistry, particularly stoichiometry[7]: "This branch of philology might be compared to the new department of *stœchiometry* in chemistry, which treats the proportions of the quantities of the elements in a state of neutralization or solution – a branch of science which everyday becomes more important […]".

[7] Stoichiometry is that branch of chemistry which deals with the relative quantities of reactants and products in chemical reactions; in this context, Lieber explicitly refers to relevant works of contemporary scholars such as Martin Heinrich Klaproth (1743-1817), Jöns Jakob Berzelius (1779-1848), and Johann Wolfgang Döbereiner (1870-1849).

With these remarks and his understanding of phonology, Lieber was much ahead of his time. Moreover, Lieber, with this short contribution, laid the foundations to make the calculation of CV proportions useful for issues of linguistic typology. This relates not only the cross-linguistic typology of different languages: with his remarks on making separate analyses for different kinds of texts (restricted, admittedly, to the rough juxtaposition of prose vs. poetry), this concerns intra-lingual specifics of text typology, as well. Both questions continue to play an important role till our days.

References

Duponceau, Peter S. (1818). English Phonology; Or, an Essay towards an Analysis and Description of the component sounds of the English Language. In: *Transactions of the American Philosophical Society, New Series, Vol. 1*; 228-264.

Grzybek, Peter (2006). A Very Early Slavic Letter Statistic and the Czech Journal 'Krok' (1841). Jan Svatopluk Presl (1791-1849). In: *Glottometrics*, 12; 88-91.

Harley, Lewis R. (1889). *Francis Lieber. His life and political philosophy*. New York: The Columbia University Press.

Isačenko, Aleksandr V. (1939/1940). Versuch einer Typologie der slavischen Sprachen. In: *Linguistica Slovaca*, 1/2; 64-76.

Kelih, Emmerich (2010a). Vokal- und Konsonantenanteil als sprachtypologisches Merkmal slawischer Literatursprachen. In: Fischer, Katrin B. et al. (eds.): *Beiträge der europäischen slavistischen Linguistik (Polyslav 13)*. München: Sagner; 70-77. [= Die Welt der Slaven Sammelbände; 40].

Kelih, Emmerich (2010b). „Wortlänge und Vokal-Konsonantenhäufigkeit: Evidenz aus slowenischen, makedonischen, tschechischen und russischen Paralleltexten." In: *Anzeiger für Slavische Philologie*, 36; 7-27.

Krámský, Jiří (1946-1948). Fonologické využití samohláskovych fonémat. In: *Linguistica Slovaca*, 4-6; 39-43.

Lieber, Francis (ed.) (1827-1832). *Encyclopædia Americana. A Popular Dictionary of Arts, Sciences, Literature, History, Politics and Biography, Brought down to the present time; including a copious collection of original articles in American biography; on the basis of the seventh edition of the German Conversations-Lexicon*. Philadelphia: Carey and Lea.

Lieber, Francis (1837). Consonants. In: *Encyclopædia Americana.* Philadelphia: Carey and Lea. Vol. III, 449-453.

Lieber, Francis (1837). On the Study of Foreign Languages, Especially of the Classic Tongues: A Letter to Hon. Albert Gallatin. In: *Southern Literary Messenger* III: 162-172. [Expanded version reprinted in: *The Miscellaneous Writings of Francis Lieber*, vol. 1. Philadelphia: J.B. Lippincott, 1881; 499-534.]

Lieber, Francis (1850). On the Vocal Sounds of Laura Bridgeman, the Blind Deaf Mute at Boston: Compared with the Elements of Phonetic Language. In: *Smithsonian Contributions to Knowledge* 2: 3-32. [Expanded version reprinted in: *The Miscellaneous Writings of Francis Lieber,* vol. 1. Philadelphia: J.B. Lippincott, 1881; 443-497.]

Lieber, Francis (1852). Plan of Thought of the American Languages. In: Schoolcraft, Henry (ed.), Historical and statistical i*nformation respecting the history, condition, and prospects of the Indian Tribes of the United States…* Vol. 2, Philadelphia: Lippincott, Grambo, 1851-1857; 46-349.

Mack, Charles R.; Lesesne, Henry H. (eds.) (2005). Francis Lieber and the Culture of the Mind: Fifteen Papers Devoted to the Life, Times, and Contributions of the Nineteenth-

century German-American Scholar, with an Excursus on Francis Lieber's Grave: Presented at the University of South Carolina's Bicentennial Year Symposium Held in Columbia, South Carolina, November 9-10, 2001. Columbia, S.C.: University of South Carolina Press.

Skalička, Vladimír (1966). Ein typologisches Konstrukt. In: *Travaux Linguistiques de Prague*, 2; 157-163.

Ycart, Bernard (2013). Letter counting: a stem cell for Cryptology, Quantitative Linguistics, and Statistics. In: *Historia Linguistica*, 40(3); 303-329. [= http://arxiv.org/abs/1211.6847]

Ycart, Bernard (2013). Alberti's letter counts. In: *Literary and Linguistic Computing*. [In print.– http://arxiv.org/abs/1210.7137]

http://archive.org/stream/encyclopaediaame03liebiala#page/n7/mode/2up