(**Note**: to appear 2006 in: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (eds.) From Data Information Analysis to Knowledge Engineering. Proceedings of the 29 th Annual Conference of the Gesellschaft für Klassifikation, 350-357. Berlin /Heidelberg: Springer**.** In case of any discrepancy with the printed version, the printed version will be the 'authorized' version.)

Cross-linguistic Computation and a Rhythm-based Classification of Languages

August Fenk and Gertraud Fenk-Oczlon

Institut für Medien- und Kommunikationswissenschaft, Universität Klagenfurt, 9020 Klagenfurt, Austria
Institut für Sprachwissenschaft und Computerlinguistik, Universität Klagenfurt, 9020 Klagenfurt, Austria

Abstract: This paper is in line with the principles of numerical taxonomy and with the program of holistic typology. It integrates the level of phonology with the morphological and syntactical level by correlating metric properties (such as n of phonemes per syllable and n of syllables per clause) with non-metric variables such as the number of morphological cases and adposition order. The study of crosslinguistic patterns of variation results in a division of languages into two main groups, depending on their rhythmical structure. Syllable-timed rhythm, as opposed to stress-timed rhythm, is closely associated with a lower complexity of syllables and a higher number of syllables per clause, with a rather high number of morphological cases and with a tendency to OV order and postpositions. These two fundamental types of language may be viewed as the "idealized" counterparts resulting from the very same and universal pattern of variation.

1 Holistic typology and numerical taxonomy

The goal of linguistic typology was from the very beginning a "classification" of languages not from the perspective of genetic and areal relations (Altmann & Lehfeldt (1973: 13)), but a "typological classification" such as the "morphological typology of the nineteenth and early twentieth centuries" (Croft (1990: 1)).
In Croft the term "classification" is used in the sense of a superordinate concept, and not, as in several other authors, as a neighbouring concept of "typology". Hempel & Oppenheim, however, suggest using "typological system" as a superordinate concept comprising "ordnende" as opposed to "klassifizierende Form" (Hempel & Oppenheim (1936: 79, 121)).
In its modern form, the domain of typology is "the study of cross-linguistic patterns of variation", says Croft (1990: 43) and attributes its earnest beginnings to Greenberg's (1966) discovery of implicational universals of morphology and word order. Greenberg's work was indeed very modern as compared with those recent studies confining themselves to seeking dependencies within syntax, within morphology, or within phonology. But his studies are, from the point of view of a "holistic typology", instances of a "partial typology". The program of a "holistic" or "systemic typology" is much older and even more ambitious with its claim to integrate also phonological properties - in addition to grammatical properties, i.e. syntactic parameters (such as word order) and morphological parameters. In the words of Georg von der Gabelentz, who introduced the term "typology" into linguistics: "Jede Sprache ist ein System, dessen sämmtliche Theile organisch zusammenhängen und zusammenwirken. /.../ Ich denke an Eigenthümlichkeiten des Wort- und des Satzbaues, an die Bevorzugung oder Verwahrlosung gewisser grammatischer Kategorien. Ich kann, ich muss mir aber auch denken, dass alles dies zugleich mit dem Lautwesen irgendwie in Wechselwirkung stehe. /.../ Aber welcher Gewinn wäre es auch, wenn wir einer Sprache auf den Kopf zusagen dürften: Du hast das und das Einzelmerkmal, folglich hast du die und die weiteren Eigenschaften und den und den Gesammtcharakter!" (von der Gabelentz (1901: 481); cited from Plank (1991:

421)). Predictivity is the goal of the "hopeful" program of holistic typology (Plank (1998)), and "numerical taxonomy" specifies the appropriate methodological principle, i.e. the principle to construct taxonomic groups with great "content of information" on the basis of "diverse character correlations in the group under study" (Sokal & Sneath (1963: 50), cited from Altmann & Lehfeldt (1973: 17)).

2 Crosslinguistic patterns found in previous studies

Our previous studies, and the present study as well, use two rather uncommon methods in order to identify crosslinguistic patterns of variation.
The first facet of this new correlational device is a "crosslinguistic" computation in the literal sense of the word: Each single language is represented by a single data pair (concerning two variables X and Y), and the computation is across the whole corpus of (a, b, c, …. n) languages.
The second facet is the use of two correlational findings as the premises from which one may infer a third correlational assumption: Given high correlations of a certain variable X with two different partners (Y, Z), this is a good hint that there might be a correlation between Y and Z as well. The higher the correlations XY and XZ, and the higher therefore the respective determination coefficients, the more plausible the inference regarding a correlation YZ. An example in the form of a syllogistic inference:

the higher Y, the lower X.
the lower X, the higher Z.

Therefore:
the higher Y, the higher Z.

"Therefore" in the conclusion means: "Therefore" it is plausible to proceed to the assumption of a positive correlation YZ. To put it more precise and more general: In the absence of any differing content-specific arguments we have to expect a positive rather than a negative sign of a third correlation in cases of equal signs in the "premises", and a negative rather than a positive sign of a third correlation in cases of different signs (+, −) in the "premises". Needless to say, that any specific expectation of this sort may prove to be wrong despite of its a priori plausibility. This way of statistical thinking is, in principle, known from the methods of partial correlation and path analysis. What seems to be new – at least within typological research – is its explicit use in order to generate new assumptions or to judge the plausibility of new assumptions respectively.
Both facets of this inferential device can best be demonstrated by means of and together with the results of our previous studies. The first one of these studies is a statistical reanalysis (Fenk-Oczlon & Fenk (1985)) of experimental data by Fenk-Oczlon (1983): In the experimental study, native speakers of 27 different languages were asked to give a written translation of a set of 22 simple declarative sentences – e.g. *The sun is shining; I thank the teacher* - and to determine the number of syllables of each of the sentences. These written translations (completely represented in the appendix of Fenk-Oczlon (1983)) allowed, moreover, to count the words per sentence and to determine the number of phonemes with the aid of grammars of the respective languages. (The results of this procedures and calculations, i.e. the characteristic values of each single language – mean n of syll./clause, mean n of words/clause, etc. - are listed up in Fenk & Fenk-Oczlon (1993, Table 4)) As expected, the language's mean number of syllables per clause was approximately in the region of Miller's (1956) magical number seven, plus or minus two. But obviously the single languages' position within this range on the continuum "n of syllables/clause" was not accidental: Dutch,

which is known for its complex syllables, encoded the semantic units with a mean of 5.05 syllables/clause; Japanese with its extremely simple syllables (or mora) marked the other end of the range with a mean of 10 syllables (or mora) per clause. We suspected the syllable-complexity (n of phonemes/syllable) being the relevant determinant. This assumption was tested by correlating the languages' mean number of syllables/clause with their mean number of phonemes/syllable.

	syll./clause	phon./syll.
Dutch	5.045	2.9732
.		
.		
English	5.772	2.6854
.		
Italian	7.500	2.1212
.		
.		
Japanese	10.227	1.8756

Table 1: The principle of a "crosslinguistic correlation" in the literal sense of the term (see correlation (a) in the text)

This was, as far as we can see, the first "crosslinguistic correlation" in the literal sense of the word, and it turned out to be highly significant (Fenk-Oczlon & Fenk (1985)):
(a) the more syllables per clause, the fewer phonemes per syllable
In a later study (Fenk & Fenk-Oczlon (1993)) with a slightly extended sample of languages we tested three further assumptions (b, c, and d). Correlation (a) indicates the view of systemic balancing effects providing a crosslinguistically "constant" or "invariant size" of simple declarative sentences. If this view holds, one has to assume a further balancing effect between word complexity (in terms of n of syllables) and the complexity of sentences (in terms of n of words):
(b) the more words per clause, the fewer syllables per word
Correlation (b) is a crosslinguistic version of Menzerath's generalization "the bigger the whole, the smaller its parts", while the following correlation (c) is a crosslinguistic version of a law actually verified by Menzerath (1954) in German. Here, the "whole" is not the sentence but the word:
(c) the fewer phonemes per syllable, the more syllables per word
Correlations (a) and (c) taken together as "premises" (see above) indicated a positive correlation (d):
(d) the more syllables per clause, the more syllables per word
The whole set of mutually dependent linear correlations (a, b, c, d) proved to be significant, and the calculations of higher-order (e.g. quadratic) functions resulted, for obvious reasons, in even higher determination coefficients. This pattern of crosslinguistic variation seems to reflect time-related constraints in sentence production and perception.

A follow-up study (Fenk-Oczlon & Fenk (1999)) with an again extended sample of now 34 languages (18 Indo-European including German, and 16 Non-Indoeuropean) could not only verify this set of correlations between metric properties but revealed, moreover, a significant association between such metric properties and the predominant word order of languages. Comparisons between Object-Verb order versus Verb-Object order and the respective t-tests significantly showed that OV order is associated with a low number of phonemes per syllable

and a high number of syllables per word and per clause, and VO order with the opposite characters. These results encouraged our search for further connections between metric and non-metric properties.

3 Connecting metric with non-metric properties

The formulation of the following hypotheses was, first of all, guided by more or less provisional ideas about interdependences between linguistic characteristics, but was assisted by the "inferential principle" described above. The linguistic arguments and the relevant chain of reasoning (for more details see Fenk-Oczlon & Fenk (2005)) resulted in a set of new hypotheses. Actually, the following list contains only 5 different correlations, because B3 is a paraphrase of A3.

A Number of morphological cases (A): a high number of cases is associated
  A1 with a low number of phonemes per syllable (r = −),
  A2 with a high humber of syllables per clause (r = +), and
  A3 with a low proportion of prepositions (r = −), i.e. a tendency to postpositions.

B Adposition order (B): a tendency to prepositions (as opposed to a tendency to postpositions), is associated
  B1 with a high number of phonemes per syllable (r = +),
  B2 with a low number of syllables per clause (r = −), and
  B3 with a low number of morphological cases (r = −).

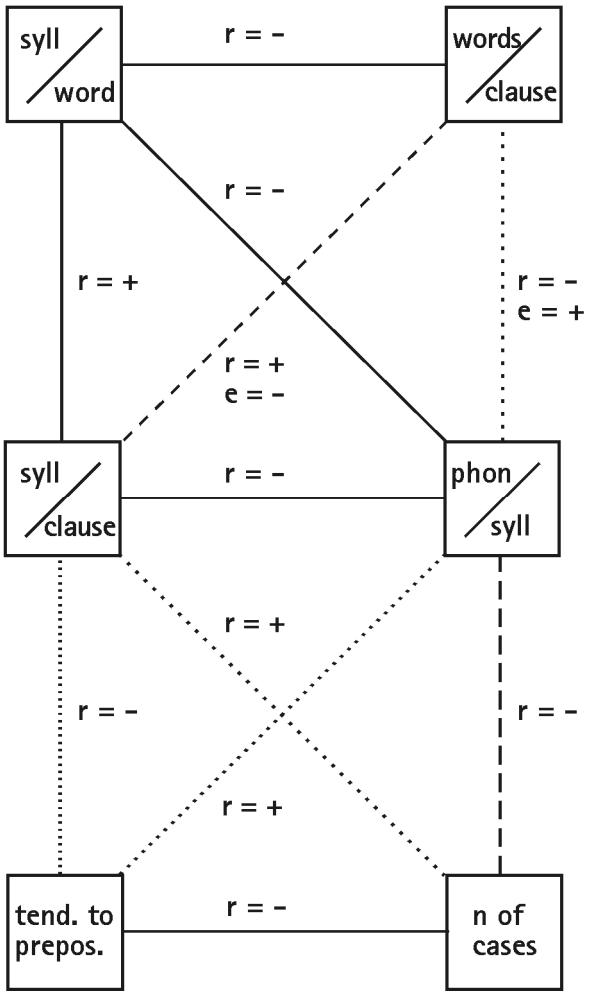

Figure 1: A correlational model connecting metric properties (in the upper part of the figure) with the two non-metric properties "tendency to prepositions" and "number of cases".
Significant correlations: solid lines
Non-significant coefficients > 0.32: broken lines
Non-significant coefficients < 0.32: dotted lines
e = expected sign differing from the sign obtained

The tendency to suffixing is generally stronger than the tendency to prefixing (e.g. Greenberg (1966)), and postpositions get more easily attached to the stem, thus forming a new semantic case (e.g. a local case). This is the linguistic argument for hypothesis A3. One might add a formal argument connecting our metric parameters with the non-metric properties A and B: Given a plausible assumption of a correlation of A or B with either "syll./clause" or "phon./syll.", this is sufficient – most apparently in the case of a "diagonal" relation in the lower part of Figure 1 – for the construction of this correlational model.

A point-biserial correlation revealed a highly significant result regarding this correlation A3: A high proportion of postpositions, or a low proportion of prepositions respectively, coincides with a high number of cases (Fenk-Oczlon & Fenk (2005)). The negative correlation B1 between the number of cases and the number of phonemes per syllable proved to be "almost significant" when calculated for only those 20 languages having case.
Figure 1 illustrates in its upper part the correlations between the metric variables and connects these complexity measures with the non-metric variables (adposition order, number of cases) in the lower part. All significant correlations correspond to the plausibility arguments explicated above. In the lower part, even the non-significant correlations correspond to those arguments. Exceptions are the two non-significant correlations in the upper part of Figure 1: Seeing the significant correlations (solid lines) of the parameter syll./word with its partners syll./clause and words/clause one should expect rather a negative sign (e = –) in a possible

correlation between these two partners, while the two significant correlations between syll./word and its "partners" words/clause and phon./syll. have the same sign and would rather suggest a positive sign (e = +) between those two partners. Actually, the result was a positive coefficient in the first case (r = + 0.328, broken diagonal line) and a negative coefficient near zero in the second case (r = − 0.013, dotted line).

4 A rhythm-based distinction between two fundamental types of language

The comparison in Table 2, though not statistically corroborated in every detail, offers a synopsis of our results so far. We should add that a high number of morphological cases (right column) will go hand in hand with separatist case exponents and a low number of morphological cases (left column) with cumulative case exponents. And it is really tempting to associate the pattern in the right column with agglutinative morphology and the pattern on the left with fusional or isolating morphology. Instead we take the speech rhythm as an anchor of typological distinction - as did Auer (1993) within phonology and Donegan & Stampe (1983) as well as Gil (1986) in the sense of a holistic approach - and as a determinant of a pattern of variation affecting phonology, morphology, and syntax. Our correlational results match the findings and interpretations of Donegan & Stampe rather than those of Gil.

All natural languages show a segmentation into intonation units, due to our breath cycle, and a segmentation of intonation units into syllables. Intonation units may be considered a special case of action units (Fenk-Oczlon & Fenk (2002)) comprising a limited number of syllables as their basic element. Smaller parts of syllables, such as vowels and consonants, are not more than "analytical devices" or "convenient fictions for use in describing speech." (Ladefoged (2001: 175)). The syllables are not only the basic elements of speech and the most appropriate crosslinguistic measure for the "size" of sentences, but represent, moreover, the single "pulses" of a language's rhythmic pattern. And this pattern is closely associated with syllable complexity: Syllable-timed rhythm with low syllable complexity (low n of phonemes per syllable), stress-timed rhythm with high syllable complexity (e.g. Roach (1982), Auer (1993), Ramus et al. (2000)). One might even argue that rhythm affects syllable complexity and that the parameter "phon./syll." in our Figure 1 is the point of impact: Changes in the rhythmic structure of a language, induced for instance by language contact, will induce changes and balancing effects in other parameters of the system. This "moving" pattern of variation, and the boundaries of variation, may be viewed as universal facts about language.

Table 2: Two fundamental types of language

stress-timed rhythm	**syllable-timed rhythm**
metric properties:	*metric properties:*
high n of phonemes per syllable	low n of phonemes per syllable
low n of syllables per clause	high n of syllables per clause
low n of syllables per word	high n of syllables per word
high n of words per clause	low n of words per clause
non-metric properties:	*non-metric properties:*
VO order	OV order
tendency to prepositions	tendency to postpositions
low n of cases	high n of cases

The two patterns figured out in Table 2 may well be considered "idealized" counterparts resulting from the very same and universal pattern of variation. Our model of this universal "groundplan" of languages includes, first of all, metric variables or otherwise quantitative variables, such as a language's number of cases. This was an advantage in constructing a correlational model of that groundplan. After integrating the data from our most recently gained translations from an English version of our test-sentences into Austronesian languages, we hope to improve the model by some kind of path analysis including, where possible, a search for the "best fitting" function between any two partners related to each other.

References:

ALTMANN, G. and LEHFELDT, W. (1973): *Allgemeine Sprachtypologie*. Wilhelm Fink, München.

AUER, P. (1993): *Is a Rhythm-based Typology Possible? A Study of the Role of Prosody in Phonological Typology.* KontRI Working Paper (University Konstanz) 21.

CROFT, W. (1990): *Typology and Universals*. Cambridge University Press, Cambridge.

DONEGAN, P. and STAMPE, D. (1983): Rhythm and the Holistic Organization of Language Structure. In: J. F. Richardson et al. (Eds.): *Papers from the Parasession on the Interplay of Phonology, Morphology and Syntax*. Chicago: CLS 1983, 337-353.

FENK, A. and FENK-OCZLON, G. (1993): Menzerath's Law and the Constant Flow of Linguistic Information. In: R. Köhler and B. Rieger (Eds.): *Contributions to Quantitative Linguistics*. Kluwer Academic Publishers, Dordrecht, 11-31.

FENK-OCZLON, G. (1983): Bedeutungseinheiten und sprachliche Segmentierung. Eine sprachvergleichende Untersuchung über kognitive Determinanten der Kernsatzlänge. Tübingen: Narr.

FENK-OCZLON, G. and FENK, A. (1985): The Mean Length of Propositions is 7 Plus Minus 2 Syllables - but the Position of Languages within this Range is not Accidental. In: G. D'Ydewalle (Ed.): *Cognition, Information Processing, and Motivation. XXIII International Congress of Psychology*. (Selected/revised papers). North-Holland, Elsevier Science Publishers B.V., Amsterdam, 355 – 359.

- (1999): Cognition, Quantitative Linguistics, and Systemic Typology. *Linguistic Typology*, 3 – 2, 151 – 177.

- (2002): The Clausal Structure of Linguistic and Pre-linguistic Behavior. In: T. Givón & B. F. Malle (Eds.): *The Evolution of Language out of Pre-Language*. (Typological Studies 53). John Benjamins, Amsterdam, 215 – 229.

- (2005): Crosslinguistic Correlations between Size of Syllables, Number of Cases, and Adposition Order. In: G. Fenk-Oczlon and Ch. Winkler (Eds.): *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler*. Gunther Narr, Tübingen, 75 – 86.

GABELENTZ, G. von der (1901): *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse.* Tauchnitz, Leipzig.

GIL, D. (1986): A Prosodic Typology of Language. *Folia Linguistica 20*, 1986, 165- 231.

GREENBERG, J. H. (1966): Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In: J-H. Greenberg (Ed.): *Universals of Language.* MIT Press, Cambridge, Mass., 73-113.

HEMPEL, C.G. and OPPENHEIM, P. (1936): *Der Typusbegriff im Lichte der neuen Logik*. A.W. Sijthoff's Uitgeversmaatschappij N.V., Leiden.

LADEFOGED, P. (2001): *Vowels and Consonants: an Introduction to the Sounds of Languages*. Blackwell Publishing, Oxford.

MENZERATH, P. (1954): *Die Architektonik des deutschen* Wortschatzes. Dümmler, Bonn.

MILLER, G.A. (1956): The Magical Number Seven, Plus or Minus Two: some Limits on our Capacity for Processing Information. *Psychological Review,* 63, 81-97.

PLANK, F. (1986): Paradigm Size, Morphological Typology, and Universal Economy. *Folia Linguistica,* 20, 29-48.

- (1991): Hypology, Typology: The Gabelentz Puzzle. *Folia Linguistica,* 25, 421 – 458.
- (1998): The Co-variation of Phonology with Morphology and Syntax: A Hopeful History. *Linguistic Typology 2*, 195-230.

RAMUS, F., HAUSER, M.D., MILLER, C., MORRIS, D., and MEHLER, J. (2000): Language Discrimination by Human Newborns and by Cotton-top Tamarin Monkeys. *Science 288*, 349-351.

ROACH, P. (1982): On the Distinction between "Stress-Timed" and "Syllable-Timed" Languages. In: D. Crystal (Ed.): *Linguistic Controversies.* Edward Arnold, London, 73-79.

SOKAL, R.R. and SNEATH, P.H.A. (1963): *Principles of Numerical Taxonomy*. W.H. Freeman, San Francisco.