The effect of morphological typology on comparative method: an attempt of quantitative assessment

Artemij Keidan
Sapienza University of Rome (artemij.keidan@uniroma1.it)

Abstract — The paper tries to give a formal answer to the following question: how did the Indo-European hypothesis become the most successful one in the comparative linguistics. I interpret this "empirical reliability" of the Indo-European as the availability of a high number of roots shared by all the involved languages (or, at least, by the most representative ones). An attempt is made to correlate the empirical reliability to the inflectionality of these languages. It is argued that, unlike morphological compositionality in agglutinative languages, the idiomaticity of derivation observed in the inflectional languages allows them to preserve more lexemes suitable for filling the cognate table. The hypothesis is tested by comparing two random groups of languages: Germanic for the inflectional type and Finno-Ugric for the agglutinative. The evidence seems to confirm the claim.

Keywords — Indo-European, Finno-Ugric, comparative method, cognates, agglutinative, inflectional, compositionality, idiomaticity

Introduction

The Indo-European (IE) family of languages, on a par perhaps only with the Semitic family, is in many ways exemplary in historical and comparative linguistics. The Indo-European genetic hypothesis appears the best proven, reliable and convincing among all such hypotheses. It is used in teaching comparative method and as a benchmark in research. It is no longer seriously questioned by anyone. To put it briefly, the IE hypothesis is distinguished by a high degree of ***empirical reliability***.

There are some obvious historical, i.e. non-linguistic, reasons for this, such as:

- the high quality of the written sources for many of the involved languages (due to long written traditions, consistent phonographic writing systems, genre diversity of the attested texts);

- the high quality of grammatical and lexicographic accounts of the involved languages;

– the high quality of the scientific effort devoted to the study of these languages (partly thanks to the early start, partly because of some obvious cultural biases).

The main claim of the present paper is that the success of the Indo-European genetic hypothesis might also have purely linguistic explanations: certain grammatical features of the IE languages apparently increase the empirical reliability of the IE genetic hypothesis. Anticipating the conclusion, such a feature should be sought in the inflectional typology shared by the oldest IE languages. Let us reword this informal claim as Hypothesis 1:

Hypothesis 1 (H_1). *The high degree of empirical reliability of the IE comparative construct is correlated with the inflectionality of its member languages.*

While mentioning inflectionality I refer principally to the ancient IE languages (such as Old Indian, Ancient Greek, Latin, Old Church Slavonic, Gothic, etc.), rather than to their later descendants that show lesser degree of inflectionality (thus, many New Indo-Aryan varieties, as well as Modern Armenian, exhibit several features of agglutinativity, while many modern Germanic languages tend towards an isolating typology).

It could be objected that such notions as “inflectional” vs. “agglutinative” are too outdated to be used as the ground for a formal analysis of linguistic facts. Rather than taking a position in this debate, I simply use them here as convenience labels denoting clusters of certain formal properties explicitly defined in the paper (see §3.1.1).

The paper has the following structure. In Section 1 basic terminology and some auxiliary concepts are defined; in Section 2, a reformulation of H_1 is proposed and two auxiliary propositions are introduced; in Section 3, the auxiliary propositions and the main hypothesis are discussed analytically and then tested empirically; Section 4 contains a brief summary of the results.

All calculations have been made using Sergey Starostin’s etymological database *The Tower of Babel* (see Starostin et al. 1998–2013), largely known by the nickname *Starling*. Accordingly, empirical data are analysed here only insofar as they are registered in *Starling*. Note that data on specific languages are presented unevenly in this resource. Thus, its IE section relies on Pokorny’s IE dictionary (Pokorny 1959), Vasmer’s Russian etymological dictionary (Vasmer 1950–1958), plus two unnamed sources for Germanic and Baltic etymology. No specific etymological dictionaries of Latin, Greek, Sanskrit etc. have been compiled into *Starling* so far. This asymmetry in the data represents a potential bias affecting all the calculations regarding the IE family as a whole, which will be consequently avoided.

Moreover, the choice of arithmetic operations involved in this work was necessarily limited to those provided by *Starling*’s search engine.

Some linguistic examples come from other sources, including dictionaries by de Vaan (2008), Holthausen (1934), and Kroonen (2013).

1 Terminology and basic notions

Unfortunately, not much has been done in order to formalise the framework and to define the ground assumptions of historical linguistics.[1] Consequently, some important notions, too often taken for granted or considered self-evident, undergo here an explicit (re)definition on a set-theoretical basis.

1.1 Comparative constructs

In the present paper I formalise the notion of "genetic hypothesis" under the label of ***comparative construct***, being an abstract construction provided with the following components:

1) a set of ***associate languages***;

2) a ***basic pattern*** of segmental correspondence;

3) a ***cognate table***.

As a working example I will use a toy comparative construct made of four associate Germanic languages (henceforth, $\mathscr{K}_{\textsc{ger}}$): Gothic, Old English, Old Norse and Old High German; only a fragment of the segmental inventory and lexicon of these languages is taken into account.[2]

1.2 Basic pattern of segmental correspondence

The easiest way to introduce the notion of basic pattern is to present it as a table. Thus, Table 1 (p. 4) displays the basic patter of segmental correspondence of the toy-construct $\mathscr{K}_{\textsc{ger}}$. Segments are presented in a normalised orthographical form, rather than in phonological transcription.

The columns of the table correspond to the associate languages of $\mathscr{K}_{\textsc{ger}}$. The rows of the table are interpreted as phonemes of the ***protolanguage*** for the given set of associate languages interpreted as ***descendants*** of the protolanguage (in the present case, Proto-Germanic). The phonemes of the protolanguage serve also as row labels. The table of the basic pattern in a real-world comparative construct presents each phoneme of each language at least once in the corresponding column; in my toy-construct, the phonologies of the associate languages are displayed only fragmentarily.

[1] I do not count as definitions such statements as "cognates [are] words that are ancestrally related" (Dekker & Zuidema 2020: 297). Some systematic attempts at defining the comparative method exist (cf. Dyen 1969, Hoenigswald 1973), but are not fully satisfying and have not gained a general support.

[2] Here, the term ***segment***/***segmental*** is used as a synonym of ***phoneme***/***phonological***.

Proto-Germanic (PG)	Gothic	Old Norse	Old English	Old High German
*a	a	a/ę/ǫ/ø	æ/a/e/ea	a/e
*ā		á/æ/ǫ́	ō/ē	ā
*ē$_1$	e		ǣ	
*ē$_2$		é	ē	ea
*eu	iu	jú/jó/ý	eo/ie	iu/io
*t	t	t	t	z/ȥȥ/ȥ
*þ	þ	þ	þ, ð	d

Table 1. Basic pattern for $\mathscr{K}_{\text{GER}}$ (based on Krahe 1960 and others)

Segments of different languages located on the same row are called ***correlatives*** of each other and ***reflexes*** of the corresponding segment of the protolanguage.

Some languages may associate several reflexes to one and the same segment of the protolanguage; in such cases we speak about a ***fork*** in the basic pattern. See e.g. the four reflexes of PG **a* in Old Norse, or the two reflexes of PG **þ* in Old English. In general, forks are supposed to be ***contextually decidable*** (in which case they are marked with "/" on Table 1).

Thus, the choice between the four reflexes of PG **a* in Old English is decided by some contextual phenomena traditionally called "mutation", "assimilation" and "breaking". With some simplification: PG **a* corresponds to OE *e* by the palatal mutation, to *a* before nasals, to *ea* by breaking (before *h*, *w*, *r* and *l*), to *æ* elsewhere. The three OHG reflexes of PG **t* are determined by the position in the word: *z* word-initially and after consonant, *ȥȥ* between vowels, *ȥ* word-finally.

Exceptionally, forks that are non-decidable on the grounds of the segmental context alone are also admitted in the basic pattern (they are marked with a comma in Table 1). Thus, PG **þ* gives two reflexes in Old (and Modern) English, which alternate in the same phonological contexts; cf. such minimal pairs as *mouth* (noun, voiceless *th*) ~ *mouth* (verb, voiced *th*), or *thy* ~ *thigh*.[3]

The opposite situation, when two different segments of the proto-language have one and the same reflex in a descendant language, is called a ***merger***. Importantly, for any pair of segments of the proto-language at least one descendant language must exist in which these two segments do not merge, i.e. have different reflexes. For example, PG *$\bar{e}_1$ and *$\bar{a}$ merge in Old Norse and OHG; PG *$\bar{e}_1$ and *$\bar{e}_2$ merge in Gothic. Old English is the only language of $\mathscr{K}_{\text{GER}}$ where both these pairs have non-merged reflexes.

[3] I do not count all the false forks resulting from orthographic irregularities in the manuscripts and inaccuracies in the descriptive grammars; thus, manuals of OHG usually give several reflexes for PG *$\bar{e}_2$ (e.g. ‹ea›, ‹ia›, ‹ie›, cfr. Ellis 1953: §3.16), but they are most likely to be interpreted as spelling variants of one and the same segment.

Considering the above, the notion of basic pattern can be understood as a function φ that maps each *contextual occurrence of a segment* in the proto-language to one and only one segment in the descendant language.[4]

The basic pattern automatically defines a set of ***particular patterns*** of segmental correspondence for each pair of associate languages, in our case: Gothic – Old Norse, Old English – Old Norse, OHG – Gothic, etc. These particular patterns may have their own forks (and, accordingly, mergers), either contextually decidable or not. For example, Gothic *e* corresponds to either *ā* or *ea* in OHG.

The classical comparative method assumes that all the patterns of segmental correspondence (either from the proto-language to a descendant, or also between descendants of the same proto-language) must satisfy the ***primitive representability condition***: a whole string from one language is mapped to a whole string in another language by mapping each segment of the first string to segments in the second string. By strings I understand words and ***formatives***.[5] It is only with this proviso that the ***segmental image*** of a string becomes a meaningful notion.[6]

For example, German *Pfad* can be considered a segmental image of English *path* because there exists a pattern of segmental correspondence φ such that

$$\varphi(path) = \varphi(p)\varphi(a)\varphi(th) = Pfad$$

1.3 Table of (root) cognates

Table 2 (p. 6) displays the cognate table associated with $\mathscr{K}_{\text{GER}}$ (the words are in the phonological form borrowed from *Starling*).

The first column shows the conventional "meaning" ascribed to the whole row (on which see below, §1.4); the remaining columns represent individual associate languages.

The cells are divided into halves. The upper half-cells contain the roots matching the language of the column and the meaning of the row. Two roots in the same row of the table are called ***root cognates***, or simply ***cognates***, of each other.[7]

[4] Ideally, the opposite should also be true (i.e., the function should be a bijection). In the real world, this is often not the case. Thus, a merger in the mapping from the proto-language to a descendant would result in a contextually non-decidable fork if we consider the reverse mapping.

[5] Formatives are understood here as pure strings of segments (I follow Polivanova 2022: Ch. 3 on this regard; cf. a somewhat similar definition in Chomsky & Halle 1968: 7–8). Consequently, unlike morphemes, formatives have no meanings of their own. Thus, the string *en* in *ox.en*, *tak.en*, *rott.en*, *strength.en*, *maid.en*, and *wood.en* corresponds to, perhaps, six different morphemes, but makes only one formative. I use periods " . " to visually separate formatives, while the terminal formative (the ***ending***) can be signalled by "=".

[6] Primitive representability condition is discussed in Polivanova (2008: 241). More technically, the pattern of segmental correspondence is expected to be a homomorphism that preserves concatenation: $\varphi(ab...n) = \varphi(a)\varphi(b)...\varphi(n)$; cf. Singh (2009: 95).

[7] Non-root cognates are also possible, but are not taken into consideration in the present paper.

The lower half-cells show the words selected as the best representatives of each root in each individual language (represented in morphonological form, i.e. with the formative structure exposed). I call such words ***providers*** of the root. The meanings of the providers are also annotated.

Meaning	Gothic	Old Norse	Old English	OHG
'barn'	*bans* *bans.t=s* 'barn'	*bās* *bās=s* 'crib'	*bōs* *bōs.ig* 'crib'	—
'game'	*laik* *laik=s* 'dance'	*leik* *leik=r* 'game'	*lāc* *lāc* 'game, match'	*leih* *leih* 'game, song'
'nest'	—	—	*nest* *nest* 'nest'	*nest* *nest* 'nest'
'dare'	*nanþ* *ana.nanþ.j=an* 'to dare'	*nenn* *nenn.in=n* 'powerful'	*nēþ* *nēþ=an* 'to dare'	*nind* *gi.nind=an* 'to dare'
'draw'	—	*aus* *aus=a* 'to draw'	—	*ōs* *ōs=en* 'to make free'
'plough'	—	—	*sulh* *sulh* 'plough'	—

Table 2. Fragment of the cognate table for the Germanic languages

More formally, the fact that root formatives R and P are cognates of each other within the comparative construct $\mathscr{K}_0$ means the following:

(i) the root formatives R and P belong to different associate languages of $\mathscr{K}_0$; let us assume that R comes from the language L and P comes from the language M;

(ii) the root formatives R and P are segmental images of each other according to the basic pattern φ of $\mathscr{K}_0$;

(iii) there is a lexeme $l(\mathsf{R})$ formed from the root R in L, and a lexeme $m(\mathsf{P})$ formed from the root P in M such that $l(\mathsf{R})$ and $m(\mathsf{P})$ are translational equivalents of each other, or, at least, present some obvious semantic similarity.

A cognate table row is called ***complete*** if all its cells are ***filled***; otherwise, the row is ***partial*** (that is, some cells are ***empty***). On Table 2 rows 2 ('game') and 4 ('dare') are complete; all the remaining rows are partial.

1.4 On the "meanings" of roots

As already mentioned, formatives (including roots) have no meanings in the proper sense of the word. Yet the construction of a cognate table assumes that the cognates located in one row are not only segmental images of each other, but also exhibit a certain semantic similarity with each other, as well as with the conventional "meaning" of the row. Now, only lexemes have meanings. Therefore, the semantic information needed to determine this conventional meaning is obtained from the providers.

Note that providers may be either ***minimal***, i.e. containing just the root plus the necessary ending (perhaps a zero), or ***non-minimal***, i.e. built by adding prefixes and/or suffixes to the root in question.

In the simplest case we obtain minimal providers being consistent translational equivalents of one another. See, for example, Old Church Slavonic, Latin, Greek, Sanskrit and Armenian cognates for PIE **dom*/*dem*/*dm* 'house' in Table 3;[8] cf. also rows 2 ('game') and 3 ('nest') in Table 2 (p. 6).

Old Irish	English	Latin	Greek	Slavonic	Lithuanian	Sanskrit	Armenian
dam	/teɪm/	*dom*	δομ	*dom*	*dim*	*dam*	*tun*
dam.n=ae 'substance'	*tame* 'to domesticate'	*dom=us* 'house'	δόμ=ος 'house'	*dom=ъ* 'house'	*dìm.st.i=s* 'porch'	*dám=as* 'house'	*tun* 'house'

Table 3. Attested cognates for the PIE root **dom*/*dem*/*dm* 'house'

Otherwise — and not so infrequently — providers, often non-minimal, are not translational equivalents of the conventional "meaning" of the root. Yet, we have no choice but using them in our cognate table. If we are lucky they are at least somewhat related to the expected meaning. E.g., the only undisputed Lithuanian provider for the series presented in the Table 3 is a derivative lexeme, whose semantic link to the notion of 'house' is acceptably close. But sometimes we have to deal with providers showing no semantic similarity at all. I call such providers ***deviant***. This may be the case of Old Irish *damnae* 'substance' and English *tame* in Table 3; cf. also the Old Norse adjective *nenninn* 'powerful' in relation to the conventional meaning of the row 4 'dare' in Table 2.[9]

1.5 Cognate table metrics

To compare different cognate tables of different comparative constructs some kind of metrics is needed. Table 4 (p. 8) sums up the ***indexes*** proposed on such regard.

The cognate table of my toy-construct $\mathscr{K}_{\text{GER}}$, displayed in Table 2 (p. 6), shows the following values: $\mathsf{Index}_{\cup} = 6$, $\mathsf{Index}_{\cap} = 2$; indexes of representation for each language are: 3 for Gothic, 4 for Old Norse and OHG, 5 for Old English; indexes of representation for roots are: 1 for 'plough', 2 for 'nest' and 'draw', 3 for 'barn', 4 for 'game' and 'dare'.

[8] Cyrillic script is transliterated except for the *yers*, which are left in the original Cyrillic shape (ъ for *ŭ* and ь for *ĭ*). Furthermore, in the Table 3 the English root is given in phonological transcription.

[9] A metaphorical justification of the deviant meaning — especially if the metaphor is paralleled in other languages of the construct — can be helpful in similar cases, but is by no means compulsory.

Index	Symbol	Explanation
Union of cognates	$\mathsf{Index}_{\cup}$	Total number of rows in the table
Intersection of cognates	$\mathsf{Index}_{\cap}$	Number of complete rows
Language representation	$\mathsf{Index}_{\updownarrow}$	Number of filled cells in a column
Root representation	$\mathsf{Index}_{\leftrightarrow}$	Number of filled cells in a row

Table 4. Cognate table metrics

Let us take into consideration the relations among the indexes in three prototypical configurations: 1) a table with no empty cells, 2) a table with few empty cells, and 3) a table with several empty cells.

1. The cognate table with no empty cells is the ***ideal*** one. In such a case, $\mathsf{Index}_{\cup} = \mathsf{Index}_{\cap}$, and both are equal to $\mathsf{Index}_{\updownarrow}$ of each language; furthermore, $\mathsf{Index}_{\leftrightarrow}$ of each root equals the number of associate languages. Ideal tables of such kind do not occur in practice.

2. In the ***optimal*** — yet actually attested — cognate tables, $\mathsf{Index}_{\updownarrow}$ and $\mathsf{Index}_{\leftrightarrow}$ vary to a little degree, respectively, from language to language and from root to root. In this case, necessarily, $\mathsf{Index}_{\cap} < \mathsf{Index}_{\cup}$. (Note that $\mathsf{Index}_{\cup}$ is always at least equal to the largest $\mathsf{Index}_{\updownarrow}$ among associate languages, in any situation.) Cognate tables for closely related languages, such as many subgroups of the IE family — e.g. Slavic, Germanic, Indo-Aryan, or the like — approximate the optimalilty.

3. The ***normal*** tables, which are the most common in practice, show significant fluctuations in both $\mathsf{Index}_{\updownarrow}$ and $\mathsf{Index}_{\leftrightarrow}$, which in turn also heavily affects the value of $\mathsf{Index}_{\cap}$. For example, the table of cognates for the IE family as a whole, as accounted for by *Starling*, while showing $\mathsf{Index}_{\cup} = 3178$, has an extremely low $\mathsf{Index}_{\cap}$, for instance, as low as eight.[10]

Such a low value of $\mathsf{Index}_{\cap}$ of $\mathscr{K}_{\mathrm{IE}}$ is justified if we remember that this construct includes many so-called *Restsprachen*, i.e. languages with an extremely low $\mathsf{Index}_{\updownarrow}$. Such languages are, e.g., Phrygian, Thracian, Ligurian, Messapic, Illyrian, Macedonian and other languages attested by scarce, fragmentary, badly preserved, and thematically uniform textual corpuses, sometimes limited to just a few words in total. By including such data in the common cognate table, the $\mathsf{Index}_{\cap}$ is caused to decrease rapidly, virtually down to zero. Therefore, in order to "optimise" a real-world cognate table, it is reasonable to exclude the *Restsprachen* from the computation, as will be done also here.

[10] The eight IE roots with the maximal $\mathsf{Index}_{\leftrightarrow}$, i.e. common to all the associate languages of the construct (at least, according to the data registered on *Starling*) are worth a mention. In semantic values: 'door', 'eat', 'know', 'name', 'stand', 'thou', 'three', 'what'.

In general, as new languages are added to the construct, the $\mathsf{Index}_{\cup}$ can only increase, while $\mathsf{Index}_{\cap}$ can only decrease. A real-world example illustrating this correlation is plotted in Figure 1. The two curves have been constructed as follows. Ten languages with the highest $\mathsf{Index}_{\updownarrow}$ have been singled out from both $\mathscr{K}_{\mathrm{IE}}$ (Indo-European comparative construct) and $\mathscr{K}_{\mathrm{UR}}$ (Uralic comparative construct). In decreasing order: Germanic, Greek, Baltic, Latin, Slavic, Sanskrit, Celtic, Avestan, Armenian, Tocharian for $\mathscr{K}_{\mathrm{IE}}$;[11] Finnish, Komi, Khanty, Saam, Estonian, Mansi, Hungarian, Udmurt, Mordovian, Mari for $\mathscr{K}_{\mathrm{UR}}$. These sets of languages have been arranged in subsets with increasing number of members (from 2 to 10, mapped on the horizontal axis), such that each additional language presented a lower $\mathsf{Index}_{\updownarrow}$; the value of $\mathsf{Index}_{\cap}$ of each of these subgroups (mapped on the vertical axis) has then been plotted on the graphic.

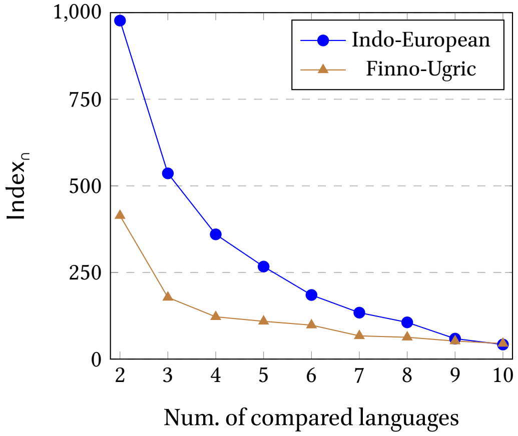

Figure 1. Germanic vs. Finno-Ugric

As Figure 1 clearly shows, whatever the starting values of $\mathsf{Index}_{\updownarrow}$ might be, for a group of languages, the value of $\mathsf{Index}_{\cap}$ near-exponentially decreases as new languages are added to the group.

1.6 Compositionality vs. idiomaticity in derivation

The question of non-minimal providers leads me to consider the phenomenon of ***derivation*** by affixation. The latter, together with ***composition***, are two basic mechanisms by

[11] To be accurate, some of these language names refer to groups rather than to single languages (e.g. Germanic, Slavic and others). However, these are the denominations used by the authors of *Starling*. As a matter of fact, the groups are usually represented by their oldest members (such as Sanskrit for Indo-Aryan, Gothic for Germanic, Old Irish for Celtic, etc.). Therefore, most likely we are still dealing with languages, rather than groups.

which new lexemes are formed in languages. Many languages — including all those under consideration here — use both such mechanisms. See some examples of both mechanisms in Table 5.

Composition	
Hungarian	*kávé.ház* ‘coffee shop’ ← *kávé* ‘coffee’ + *ház* ‘house’ *adó.mentes* ‘tax-free’ ← *adó* ‘tax’ + *mentes* ‘free’ *jég.hideg* ‘freezing’ ← *jég* ‘ice’ + *hideg* ‘cold’
Finnish	*englannin.torvi* ‘English horn’ ← *englannin* ‘English’ + *torvi* ‘horn’ *adessiivi.sija* ‘adessive case’ ← *adessiivi* ‘adessive’ + *sija* ‘place’ *eristys.nauha* ‘insulating tape’ ← *eristys* ‘insulation’ + *nauha* ‘tape’
Gothic	*nahta.mats* ‘evening meal’ ← *nahts* ‘night’ + *mats* ‘meal’ *filu.waurdjan* ‘be talkative’ ← *filu* ‘much’ + *waurdjan* ‘talk’ *mikil.þuhts* ‘arrogant’ ← *mikils* ‘great’ + *þuhtus* ‘mind’ *hrainja.hairts* ‘pure-hearted’ ← *hrainjis* ‘pure’ + *hairto* ‘heart’
German	*Zeitungs.anzeige* ‘newspaper ad’ ← *Zeitung* ‘newspaper’ + *Anzeige* ‘ad’ *Fahr.stil* ‘driving style’ ← *Fahr* ‘driving’ + *Stil* ‘style’ *sitzen.bleiben* ‘repeat the grade’ ← *sitzen* ‘seat’ + *bleiben* ‘remain’

Derivation	
Hungarian	*temet.és* ‘funeral’ ← *temet* ‘bury’ *alázatos.ság* ‘humility’ ← *alázatos* ‘humble’ *árt.atlan* ‘harmless’ ← *árt* ‘harm’
Finnish	*aasia.lainen* ‘Asiatic’ ← *Aasia* ‘Asia’ *huole.llinen* ‘careful’ ← *huoli* ‘care’ *ost.ella* ‘to shop’ ← *ost.aa* ‘to buy’
Gothic	*fra.wairpan* ‘throw away’ ← *wairpan* ‘throw’ *andbaht.i* ‘service’ ← *andbahts* ‘servant’ *wiþra.gaggan* ‘go towards’ ← *gaggan* ‘go’ *gait.ein* ‘young goat’ ← *gaits* ‘goat’
German	*ver.standen* ‘understand’ ← *standen* ‘stand’ *zweifel.haft* ‘doubtful’ ← *Zweifel* ‘doubt’ *Ge.danke* ‘thought’ ← *denken* ‘think’

Table 5. Examples of composition and derivation by affixation

Composition consists in conjoining two elements of the “word” rank, called ***components***, into a larger unity called ***compound***, still of the same rank.[12]

In derivation we deal with ***derivational pairs***, i.e. pairs of lexemes whereby a ***derivative*** lexeme is derived by affixation from the ***simple***. The derivative contains all the formatives of the simple, plus some additional formatives (affixes). Examples of deriva-

[12] Ranks correspond to “pieces” of varying length that an utterance is divided in, from rank 1 “segment”, to rank 2 “formative”, rank 3 “word”, etc. Entities that are substantially different may still belong to one the same rank; thus, the rank “word” comprises such things as *word-forms*, *stems*, *words*, and also *compounds*.

tional pairs include: *write – writ.er*, *pig – pig.let*, *employ – employ.ee*, *nation – nation.al*, etc. A collection of similarly built derivational pairs forms a ***derivational type***. In theory, all the members of a derivational type share one and the same ***derivational formula***. Consider the derivational type «deverbal nouns suffixed with *er*/*or*/*ar*» in English: *writ.er*, *runn.er*, *sail.or*, *begg.ar*, etc. This type can be distilled to a formula like this: «X.*er* is the person Y who performs the action X».

The semantic opacity/transparency of the lexemes built by either of the two mechanisms is now under scrutiny. Is the relationship between the meaning of the compound and the meanings of its components semantically transparent or opaque? Is the relationship between the meaning of the derivative and the corresponding simple word, plus the added affix semantically transparent or opaque?

Unsurprisingly, I will call the result of a word-forming process ***compositional*** if its meaning is a composition of the meanings of its parts; otherwise, it will be called ***idiomatic***.

As a general principle, I assume that the compositionality usually prevails in compounds, while the idiomaticity occurs more easily in derivation by affixation, at least in the Indo-European languages. However, counterexamples are also well-known.

The examples of compositionality (semantic transparency) in compounds are everywhere, cf.: English *book.store*, *head.ache*, or *corn.bread*, German *Fahr.stil* 'driving-style', Gothic *nahta.mats* 'evening meal', Finnish *englannin.torvi* 'English horn', Hungarian *adó.mentes* 'tax-free'; see more examples on Table 5 (p. 10).

Yet, compounds exhibiting some idiomaticity (semantic opacity) with respect to the meanings of their components are also well-known, even if they occur less frequently, cf.: English *butter.fly* (not a fly, no relation to butter), *hedge.hog* (not a hog, no relation to a hedge), or *green.mail* (neither mail, nor green), German *sitzen.bleiben* 'repeat a grade' (rather than, literally, 'remain seated'), or Gothic *mikil.þuhts* 'arrogant' (rather than, literally, 'large-minded').

Contrariwise, in derivatives the semantic link between the parts becomes opaque more easily. There are myriad examples of idiomaticity in derivation. For example, in the following German words, albeit all derived by adding the prefix *ver* to the base, conserve almost no semantic relation with the simple, neither do they share any specific "meaning element" despite the common prefix: *ver.stehen* 'understand' from *stehen* 'to stand', *ver.legen* 'embarrassed' from *legen* 'to lay', *ver.geben.s* 'in vain' from *geben* 'to give'.

This does not exclude the fact that a higher degree of compositionality is sometimes attested in affixation; cf. the German suffix *heit* regularly used for forming abstract nouns from quality adjectives: *Leicht.heit* 'lightness' from *leicht* 'easy', *Schön.heit* 'beauty' from *schön* 'beautiful', *Neu.heit* 'novelty' from *neu* 'new'.

The derivational type of English deverbal nouns in *er*/*or*/*ar* mentioned above is a good example of idiomaticity. Its derivational formula describes the meaning of the derivative quite accurately in many, but not all, cases. Thus, a *work.er* is effectively someone who works, a *rule break.er* is someone who breaks rules, a *read.er* is someone who reads, etc. However, in many other derivatives the formula does not apply. For example, the base of derivation may be a noun, rather than a verb; in such cases, the derivative may refer to a person's quality (as in *boom.er*), profession (as in *astrolog.er*), geographical origin (as in *Boston.er*), and so on. Or, alternatively, the derivative may indicate some other valency of the original verb, than the performer of the action, as in *disclaim.er* 'the result of disclaiming', *mix.er* 'the instrument of mixing', *merg.er* 'the process of merging', or *din.er* 'the place where people dine'. We may be tempted to produce a specific formula for each group of these exceptions, but this eventually leads to a complete atomisation of the grammatical description: if we applied this method consistently, we would need as many formulae as there are derivatives.[13]

Note also that meanings do not add up in the same elementary way as do the natural numbers or line segments. In compounds, the meaning of the whole is inferred, perhaps in some non-trivial manner, from the meanings of its parts.[14] The same applies even more so in derivation, where the meaning of the derivative lexeme is "calculated" from the simple, plus the new formative(s), according to a highly abstract derivational formula.

Consequently, this brings to the situation whereby compounds and derivative lexemes, although idiomatic to a certain degree, may nevertheless preserves some compositionality. Therefore, the evaluation of the compositionality/idiomaticity of a compound or a derivative requires not a binary, but rather a continuous scale, from 100% compositional to 100% idiomatic, admitting all the values in-between.[15] How exactly these values are calculated is another question. For illustrative purposes only I present Table 6 (p. 13) showing five English compounds ordered on such a continuous scale (from maximum "Comp." to maximum "Idiom."); the figures are totally subjective and hand-weaven.

[13] Cf. Bauer (2001: 203): "[...] there is no absolute answer to the question of how many different *-er* affixes there are in English, there are only different possible analyses of the data".

[14] For a contemporary account of this fact see, e.g., Partee (1984) and Cruse (2000: Section 4). Old Indian grammarians suggested a classification of compounds based on how the meanings are combined: in the *tatpuruṣa* type the meaning of the whole is constructed as a *dependency* of one part from the other (cf. *handshake* being the shake of a hand); in the *dvandva* type the whole meaning is a *conjunction* of two component meanings on par (as in *singer-songwriter*, i.e. someone who is both a singer and a songwriter at the same time); in the *bahuvrihi* type the whole meaning is derived by one of the preceding formulae, but it is then used as an epithet of some external referent not mentioned among the compound's members (as in *butterfingers*, said about someone who drops things easily).

[15] Note that the debate on compositionality in derivation has often adopted a purely dualistic approach: compositionality is either presented as an exceptionless principle, or denied entirely; cfr. on this Bertinetto (1995). However, gradualist views are also attested, see Hoeksema (2000: §2.3).

Word	Comp.	Idiom.
daytime	100%	0%
grandmother	70%	30%
greenhouse	50%	50%
lighthouse	30%	70%
ladybird	0%	100%

Table 6. Examples of compounds varying by the Comp./Idiom. value.

In case we need to evaluate the idiomaticity/compositionality of a whole system of derivatives (or compounds), the average value for all such lexemes may be used. In this regard, observe that in languages of the "traditional" Indo-European type, derivation by affixation shows a general tendency towards idiomaticity. The grammars of such languages *suggest* the most likely meaning of the derivative, even if nothing *guarantees* that one could always guess it correctly. For example, certain idiomatic referential restrictions may occur; thus, a *writ.er* is clearly someone who writes; however, not anyone who writes something, but only someone who produces written material professionally. The extreme outcome of this tendency towards idiomaticity are the famous ***wildcard-affixes***, that are widely used in IE languages as word-forming mechanisms, while showing no distinct semantic value.[16]

On the contrary, the system of derivation by affixation in Uralic (Finno-Ugric) languages shows a clear tendency toward compositionality. If we consider verbal and nominal suffixes in Finnish or Hungarian, we conclude that the meanings of the derivatives are easily predictable, the referential restrictions are few, and the wildcard-affixes are rare, if any.

2 Auxiliary propositions

2.1 A new formulation of the main hypothesis

Let us return to the notion of *empirical reliability*, mentioned in the main hypothesis of the present paper (H_1). In order not to appear too arbitrary, this hypothesis should be substantiated with a data-oriented interpretation. I assume without justification that such an interpretation must be expressed in terms of cognate tables metrics, i.e. by the means of the indexes defined in §1.5. The challenge now is to determine which of those indexes is the best proxy of the empirical reliability.

[16] A famous example of a wildcard-affix is the IE "velar suffix", abundantly attested in many ancient and modern IE languages (see Ciancaglini 2012); already Old Indian grammarians considered it *anartha* 'meaningless'.

At a very general level, the more roots there are in the table, the happier the linguist. Or, conversely, the fewer empty cells, the better. Which index would be the best proxy of the "density" of the empty cells in a cognate table?

The present paper assumes that the number of empty cells is inversely correlated to the Index$_{\cap}$ of the cognate table. That amounts to say: the greater the number of roots common to all the associate languages of a comparative construct, the higher its empirical reliability.

In fact, the value of Index$_{\cup}$ (i.e., number of rows) cannot be used for such a purpose, since adding new rows does not necessarily prevent empty cells from appearing. We can easily obtain a very "tall", yet significantly underpopulated, cognate table.

Neither is the number of languages (i.e., the number of columns) correlated to the density of the table, since adding new languages (columns) cannot but increase the number of empty cells.

This allows me to reformulate H_1 as H_2:

Hypothesis 2 (H_2). *The Indo-European comparative construct exhibits a high* Index$_{\cap}$*, given the inflectionality of its associate languages.*

Using Index$_{\cap}$ as a measure of empirical reliability is also a realistic representation of what comparative linguists do. Indeed, this same principle — even if differently formulated — has been proposed by others as a quality criterion in comparative linguistics.[17]

Accordingly, the paucity of common roots is often cited as a major flaw of certain debated comparative constructs.[18]

Note that the other indexes can indirectly affect the value of Index$_{\cap}$ (see §1.5). Therefore, for a meaningful comparison between different constructs, it is necessary to start from equal conditions: the same number of associate languages provided with similar ranges of Index$_{\updownarrow}$ values.

2.2 Additional theses and inference schema

In my argumentation I will consider two additional theses; the first resumes the informal predictions from §1.6.

[17] See the following claim in a methodological paper by Doerfer (1981: §3.1): "[v]ery important is the distribution of common terms [...]. Let us assume there are three languages to be investigated: A, B, C. Then we may presume genetic relationship under these conditions: 1) When many general comparisons [...] can be found, i.e., comparable roots in all the three languages A, B, C [...]."

[18] Such is the case with the Altaic family: notwithstanding a large collection of — presumed — Altaic etymologies, its supporters are blamed for not being able to provide a reasonable corpus of roots common to all the assumed branches of this family, see Vovin (2005) and Ciancaglini (2009). According to the data from *Starling*, the Index$_{\cap}$ for the Altaic languages is just 321, which represents a drammatic decrease considering how large is Index$_{\cup}$ (around 2800 roots) and how few languages are associate to this construct, in comparison with e.g. the Indo-European construct (namely, 5 vs. 14).

Thesis A ($\mathbf{Th_A}$: inflectionality ⇒ idiomaticity). *Inflectional languages show idiomatic derivation, while agglutinative languages show compositional derivation.*

Thesis B ($\mathbf{Th_B}$: idiomaticity ⇒ larger $\mathsf{Index}_\cap$). *Other conditions being equal, a set of languages with idiomatic derivation shows a larger value of* $\mathsf{Index}_\cap$ *than a set with compositional derivation.*

Let us present the proof scheme of the present paper in advance. As a preliminary result observe that, by the transitivity of implication, from the truth (or plausibility) of $\mathrm{Th_A}$ and $\mathrm{Th_B}$ logically follows the truth (or plausibility) of the conclusion $\mathrm{Th_C}$:

Thesis C ($\mathbf{Th_C}$: inflectionality ⇒ larger $\mathsf{Index}_\cap$). *Other conditions being equal, a set of inflectional languages shows a larger value of* $\mathsf{Index}_\cap$ *than a set of agglutinative languages.*

Since the ancient Indo-European languages are mostly inflectional, from $\mathrm{Th_C}$ follows the truth of $\mathrm{H_2}$. In turn, thanks to the agreement formulated in §2.1, $\mathrm{H_2}$ implies also the truth of $\mathrm{H_1}$, namely the present paper's initial hypothesis.

Now all that remains is to prove the truth (or, at least, plausibility) of $\mathrm{Th_A}$ and $\mathrm{Th_B}$. Such proofs are provided in Section 3.

3 Arguments and proofs

3.1 $\mathrm{Th_A}$: inflectionality ⇒ idiomaticity

For any language L,
if [L is inflectional]
then [L has idiomatic derivation]

Note that the inflectionality of a language is essentially a gradual, rather than binary, property: some languages are *more* inflectional, others are *less*. For example, Latin is more inflectional than Romanian, and Romanian is more inflectional than French; Gothic is more inflectional than German, German is more inflectional than Swedish, and Swedish is more inflectional than English.

Other properties discussed here are also gradual. Accordingly, regardless of how we measure the "degree of inflectionality", the "degree of agglutination", the "degree of idiomaticity of derivation" and the "degree of compositionality of derivation", $\mathrm{Th_A}$ can be reformulated in a somewhat relaxed form:

Thesis A′ ($\mathbf{Th_{A'}}$: inflectionality ⇒ idiomaticity). *The more inflectional a language, the more idiomatic derivation it shows; the more agglutinative a language, the more compositional derivation it shows.*

An analytic justification of $\mathrm{Th}_{A'}$ will be presented in §3.1.1; empirical evidence will be presented in §3.1.2.

3.1.1 Analytic justification of $\mathrm{Th}_{A'}$

On terminology. In order to draw a distinction between inflectional and agglutinative languages I use a set of diagnostic features which are similar to those used for this purpose elsewhere in the literature (e.g. Plungjan 2001). I recognise that, nowadays, this classical typological dichotomy is generally considered untenable. Testing of larger samples of languages proved that these features simply do not cluster together: too many languages show inflectional behaviour in certain aspects, but equally convincing agglutinative behaviour in other aspects. Each specific feature is still a valid measure of linguistic diversity, but no significant evidence of correlation between such features seems attested cross-linguistically.[19]

I nevertheless keep using *inflectional* and *agglutinative* as cover terms for the clusters of the parameter values used in my argument. Thus I use the following diagnostic properties: I) clarity of formative boundaries, II) functional consistency, III) morphological freedom.

Clarity of boundaries. I call agglutinative those languages where formative boundaries are clearly detectable. Contrariwise, in inflectional languages formative boundaries undergo a more or less deep segmental processing, so that they are often obscured, or fused altogether, whence such languages are also called *fusional*. This is obviously a gradual property.

The maximal clarity corresponds to no processing at all at the formative boundaries (see Table 8, Page 20, line I, label "→1") . Languages considered as exemplary agglutinative, both traditionally and in this paper, present clear boundaries. Thus, a complete absence of fusion characterises Old Japanese, while Modern Japanese shows some signs of fusion. In the Japanese examples below, the boundaries between the verb base and the gerund marker are processed by special transformation rules, while in their Old Japanese antecedents there were simply no transformations whatsoever; see Frellesvig (2010: 193). Here, and hereinafter, the segmental output of the formative joining, with or without segmental processing, is highlighted in bold face.

- OJ *shin.u+te* → *shi**nite*** 'dying'; cf. MJ *shi**nde***;

- OJ *ok.u+te* → *o**kite*** 'putting'; cf. MJ *o**ite***;

[19] Haspelmath (2009) concludes, perhaps too drastically, that what he calls the "agglutination hypothesis" has been proven wrong.

- OJ *asob.u+te* → *aso**bit**e* ‘playing’; cf. MJ *aso**nd**e*.

In Hungarian, boundaries between formatives are usually not subject to any segmental processing; some cases of slightly obscured boundaries are observed almost exclusively in inflection, see Rounds (2001: 29; 213–214).

- clear boundaries in derivation: *kér+vény* → *kér**v**ény* ‘questionnaire’, *kiad + vány* → *kia**dv**ány* ‘publication’; cf. a certain obscurity in *erdő+ész* → *er**dé**sz* ‘forester’;
- clear boundaries in inflection, for instance in the form of Pres1Pl: *ad+juk* → ***adj**uk* ‘we give’, *kér+jük* → *ké**rj**ük* ‘we ask’; cf. obscured boundaries in: *olvas + juk* → *olva**ss**uk* ‘we read’, *főz+jük* → *fő**zz**ük* ‘we look at’.

The minimal clarity, i.e. the maximal obscurity, of boundaries between formatives (see Table 8, line I, label “→0”) is attested, for example, when two neighbouring segments at a formative boundary are substituted by a third one, as happens with such forms as Greek φυλάσσω or Sanskrit *ūḍ*^{h}*á* among the following examples.

- The Greek present tense suffix -jω merges with the last segment of the root with a very complicated processing, e.g. φυλακ+jω → φυλάσσω ‘I guard’, κοπ+jω → κόπτω ‘I cut’, βαν+jω → βαίνω ‘I go’, ἐλπιδ+jω → ἐλπίζω ‘I hope’, ἀγγελ+jω → ἀγγέ**λλ**ω ‘announce’.
- The Sanskrit Past participle suffix *ta* likewise undergoes a heavy segmental processing while merging with the last segment of the root, e.g. *uh+ta* → ***ūḍ***^{h}*a* ‘brought’, *ij+ta* → *i**ṣṭ**a* ‘sacrificed’; *b*^{h}*aj+ta* → *b*^{h}*a**kt**a* ‘divided’, *gm+ta* → ***g**a**t**a* ‘gone’.
- The Latin prefix *ad* merges with the initial consonant of the root up to complete assimilation: *ad+curro* → *a**cc**urro* ‘I am in a hurry to help’; *ad+genero* → *a**gg**enero* ‘I generate’, *ad+laboro* → *a**ll**aboro* ‘I work with effort’, *ad+fuere* → *a**rf**uere* ‘they attended’ (archaic).

Functional consistency. I consider agglutinative those languages in which any formative points to exactly one grammatical function, and conversely, in which there is exactly one formative per function. I call this ***one-to-one configuration***. On the contrary, inflectional languages present various deviations from the one-to-one principle:

- many-to-one configuration, as in the so-called ***cumulative affixes***, such as Latin adjectival ending *us* marking three categorial values: masculine (gender), singular (number), and nominative (case);

- one-to-many configuration, as in ***non-contextual allomorphy***, when one grammatical value is associated to multiple formatives, e.g. English past participle expressed by either *ed* or *en* (among others);

- one-to-zero configuration, as in the so-called ***zero endings***, cf. Latin NomSg forms *puer=∅* 'child' vs. *lup=us* 'wolf';

- zero-to-one configuration, as in ***null affixes***, i.e. formatives that are clearly visible in a word but have no distinct semantic or grammatical value, cf. English *for* in *for.get*, *for.bid*, *for.give*.

Functional consistency is also a gradual feature; only for convenience can languages be considered totally functionally consistent (see Table 8, page 20, line II, label "→1") or totally functionally ambiguous (see Table 8, line II, label "→0"). Real languages, while being functionally consistent as a general principle, may occasionally show functional ambiguity (and vice versa). Thus, many Balto-Finnic languages, while generally respecting functional consistency (one-to-one principle), present occasional violations of this principle, especially under the form of one-to-zero and many-to-one configurations. Consider this fragment of the paradigm of two Estonian nouns, *küla* 'village' and *algus* 'beginning' displayed on Table 7.

	NomSg	GenSg	AccSg	PartSg
algus 'beginning'	*algus=∅*	*algus=e*	*algus=e*	*algus=t*
küla 'village'	*küla=∅*	*küla=∅*	*küla=∅*	*küla=∅*

Table 7. Estonian case-endings

Here, one and the same formative is used for marking multiple cases (the ending *e* in genitive and accusative). The zero ending also presents rich evidence, particularly when the zero marks multiple case-values. Note that all the remaining case-endings, which are not displayed on Table 7, do fully respect the one-to-one principle.

Morphological freedom. The less unpredictable the restrictions that a language puts on the type of bases to which an affix can be attached, the more morphologically free it is. The more selective a language is in this regard, the less morphologically free it is. Different degrees of morphological selectivity are possible. The most obvious kind of selectivity differentiates affixes by the lexical class of the base (e.g. noun-only affixes vs. verb-only affixes). But the most relevant cases are those in which functionally equivalent endings are selected by different groups of lexemes within the same lexical class.

The notion of functional equivalency requires additional explanation. Let us assume a deeper and more fundamental property of a language: its ***paradigmaticity***. A prototypically paradigmatic morphological system shows the following properties.

- Each ***lexical class*** has its own set of ***grammatical categories***.
- For each lexeme there is a set of word-forms that are organised into a ***paradigmatic table***, whose shape and dimensionality is imposed by the grammatical categories of the lexical class to which this lexeme belongs.
- In each word-form we can always individuate its ***ending*** — usually, the terminal formative — and detect the categorial values that classify this ending, i.e. to retrieve the ***paradigmatic address*** of this word-form in the paradigmatic table.
- There must be exactly one ending per each cell of the paradigmatic table.
- The same values that are used for classifying word-forms in the paradigmatic table, are also used to formulate syntactic requests in the sentence (as in the case of verbal government of nouns).[20]

Accordingly, two (or more) formatives that happen to have the same paradigmatic address or, equivalently, are requested in the same syntactical contexts, are considered ***functionally equivalent***. In such cases, if the choice between these formatives is only determined by the ***paradigmatic type*** of the base (i.e., to what declension or conjugation the base belongs), and cannot be semantically or phonologically predicted, then the language shows the maximal degree of morphological selectivity (minimal freedom).

Thus, Latin formatives *ant* and *unt* are functionally equivalent as long as they both have the same paradigmatic address (indicative, active, present, 3rd, plural), appear in the same syntactic contexts (e.g., in agreement with a plural 3rd person subject), but their distribution depends on an unpredictable lexical property of the base; namely, whether it belongs to the 1st or the 3rd conjugation. I consider the maximal morphological selectivity a symptom of inflectionality (see Table 8, page 20, line III, label "→0").

Contrariwise, in the absence of paradigmaticity, the notion of paradigmatic address becomes meaningless, so that no pair of formatives can ever have the same paradigmatic address, therefore, there is no functional equivalency and, as a consequence, no selection among equivalent formatives occurs. In languages of this type all grammatical marks are different from each other, and each of them occurs only to express some semantic, pragmatic or stylistic value desired by the speaker, rather than to meet the specific syntactic requirements of the context. No two marks can be ever said to be "the same thing".

[20] For a more analytical presentation of this approach see Polivanova (2022: Ch. 3–6).

One such case is represented by Korean, whose grammatical "particles" have no restrictions on the base they can attach to and very few restrictions concerning the syntactic contexts they occur in. There are even "particles" that are compatible with both nominal and verbal bases.[21] I consider this major morphological freedom a symptom of agglutinativity (see Table 8, page 20, line III, label "→1").

Morphological freedom vs. selectivity is a gradual property, as is confirmed by such "intermediate" languages as Hungarian. Although Hungarian does not impose lexical restrictions on affixes as a general rule (e.g., it has no varying declensions for nouns), it nevertheless may exceptionally exhibit synonymic affixes. However, unlike strictly paradigmatic languages, the selection between synonymic suffixes, besides being lexically triggered, also conveys additional semantic information. Thus, Hungarian has two "conjugations": simple verbs and *ik*-verbs (Rounds 2001: §4.1.1). The verbs of the *ik*-conjugation tend to be markedly intransitive, which is especially evident when both forms can be formed from one base, cf. such pair of verbs as *tör*=∅ 's/he breaks something' vs. *tör*=*ik* 's/he breaks (intr.)'. However, many other such pairs are semantically unrelated, e.g.: *ér*=∅ 's/he arrives' vs. *ér*=*ik* 's/he ripens'.[22]

Interim recap. The results of the discussion above are summed up in Table 8. The totality of "→0" marks (i.e. "close to 0") by each of the three parameters, labelled as *inflectional*, is precisely the quality that prevents the derivation from being compositional, i.e. increases its idiomaticity. On the contrary, the totality of "→1" marks (i.e. "close to 1"), labelled as *agglutinative*, is the quality that assures the compositionality of derivation.

	Parameter	Value	
		Agglutinative	Inflectional
I.	Boundaries' clarity (vs. obscurity)	→1	→0
II.	Functional consistency (vs. ambiguity)	→1	→0
III.	Morphological freedom (vs. selectivity)	→1	→0

Table 8

Indeed, if a derivational affix can mean many different things, or even nothing at all, then the derivational formula for a derivative, and therefore its meaning, become extremely hard to predict and even to account for; the more so if such an affix has indefinite boundaries and is subject to random selectional conditions. On the other hand,

[21] Apparently, only a few particles, termed *keys* by Brečalova (2009), are strictly imposed by the syntactic context. Note that the small amount of allomorphy that is attested Korean is always phonologically determined and does not involve any sort of morphological selection.

[22] Although Balto-Finnic languages may exhibit different "declensions" and "conjugations" (cf. the Estonian example on Table 7, page 18), the real distinction there does not reside in the endings per se.

if formatives have unique non-empty functions, their boundaries are always detectable and there are no lexical constraints on their selection, then the derivational formulae are easily retrievable, and the meanings of the derivatives are calculated with less difficulty.

In other words, the defining features of the agglutinative vs. inflectional distinction are also the causing factors of compositionality vs. idiomaticity in derivation.[23]

3.1.2 Evidence

Here is a selection of examples of idiomaticity in inflectional languages and compositionality in agglutinative languages. To begin with, let us consider the idiomatic nature of derivation in Old Church Slavonic and Gothic, both representing a high degree of inflectionality in the sense defined here.

Table 9 shows derivatives from two Old Church Slavonic roots. Two signs of idiomaticity are observable here: 1) some words are formally derivatives — they are built from a root plus some affixes — but have a non-complex meaning (e.g. *vrěmę* 'time', or *oblakъ* 'cloud'); 2) same-root derivatives refer to semantic fields completely unrelated to each other (compare such pairs as 'time' and 'gateway', or 'cloud' and 'wolf'[24]).

Root (all allomorphs)	Derivative lexemes
vrat/vrašt/vrět/vrьt	*vrěmę* (*vrět.mę*) 'time' *vъz.vrašt.enije* 'return' *vrьt.ěti* 'to turn' *vratъ* 'gateway'
vlak/vlač/vlьk/vlьč/vlěk	*izvlěšti* (*iz.vlěk.ti*) 'extract' *oblakъ* (*ob.vlak.ъ*) 'cloud' *oblačiti* (*ob.vlač.iti*) 'to dress' *vlьkъ* 'wolf' *vlьčьcь* 'thistle'

Table 9. Old Church Slavonic derivatives; based on Polivanova (2013)

[23] Cf. similar conclusions in Plungjan (2001: 676): "In flective languages [...] [d]erivational markers [...] are generally non-productive and idiomatic. [...] By contrast, derivational markers found in agglutinative languages are, as a rule, productive [...]". His "productive" means approximatively the same as my "compositional".

[24] A Slavicist may dissent from the etymological connection of *vlьkъ* 'wolf' (and hence of *vlьčьcь* 'thistle', lit. 'little wolf') to the other lexemes listed in this example, all connected to the idea of 'dragging'. This criticism might not be baseless diachronically. However, within the synchrony of Old Church Slavonic, there are no strong indications against keeping these lexemes together. This etymological hypothesis is mentioned by Vasmer (1950–1958) and fully accepted by Polivanova (2013).

In the Gothic examples on Table 10, two prefixes are considered. The verbs that bear these prefixes are completely idiomatic, in the sense that they are not associated to any semantic constant, according to the common prefix. The simple and the derivative verbs may not differ at all in meaning (e.g. the pair *greipan* – *fairgreipan*, both 'seize'), or may be semantically inconsistent (e.g. the pair *sitan* 'sit' – *dissitan* 'seize').

Prefix	Simple verb	Derived verb
fair	*aihan* 'have'	*fair.aihan* 'take part'
	greipan 'seize'	*fair.greipan* 'seize'
	waurkjan 'work'	*fair.waurkjan* 'gain'
dis	*sitan* 'sit'	*dis.sitan* 'seize'
	haban 'have'	*dis.haban* 'overcome'
	sigqan 'sink'	*dis.sigqan* 'sink' (about the sun)

Table 10. Gothic derivatives; based on Bucsko (2011)

Let us now consider derivation in Finnish and Hungarian, both representing a good, although not maximal, degree of agglutinativity in the sense defined here.

In Finnish, derivation by suffixation is quite common. Suffixes are very productive, they can line up in long chains, and can even be repeated, especially if these are verb-related suffixes (markers of causative, frequentative, reflexive, etc.). Segmental variation of suffixes is only admitted as a contextually-decidable allomorphy, rather than by lexical restrictions on affix selection.

Table 11 displays a set of Finnish simple lexemes together with suffixed derivatives thereof: semantics is clearly compositional between the pairs, and the semantic contribution of suffixes is also constant across lexemes.

Suffix	Simple noun	Derivative noun
sto/*stö*	*kirja* 'book'	*kirja.sto* 'library'
	laiva 'ship'	*laiva.sto* 'fleet'
	nimi 'name'	*nimi.stö* 'nomenclature'
llinen	*kirja* 'book'	*kirja.llinen* 'bookish'
	yö 'night'	*yö.llinen* 'nightly'
	hetki 'moment'	*hetke.llinen* 'momentary'
la/*lä*	*Kaleva* (proper name)	*Kaleva.la* 'land of Kaleva'
	ravinto 'meal'	*ravinto.la* 'restaurant'
	kana 'hen'	*kana.la* 'henhouse'
	setä 'uncle'	*setä.lä* 'uncle's home'
	kylpy 'bath'	*kylpy.lä* 'bathing place'

Table 11. Finnish derivatives; based on Karlsson (2018) and Hakulinen (1961)

I also give some examples of compositionality from Hungarian. Table 12 shows a series of derivatives formed by adding two suffixes, individually or sequentially, each with a constant derivative function (thus, *ász/ész* suffix forms the *nomina agentis*, while the *at/et* forms the *nomina actionis*).

Simple lexeme	*ász/ész* derivative	*at/et* derivative
épít 'building'	*épít.ész* 'architect'	*épít.ész.et* 'architecture'
mű 'artificial'	*műv.ész* 'artist'	*műv.ész.et* 'art'
szobor 'statue'	*szobr.ász* 'sculptor'	*szobr.ász.at* 'sculpture'
nyelv 'language'	*nyelv.ész* 'linguist'	*nyelv.ész.et* 'linguistics'
felel 'to answer'		*felel.et* 'answer'
él 'to live'		*él.et* 'life'

Table 12. Hungarian derivatives; based on Rounds (2001)

Many other examples showing the idiomatic nature of derivation in inflectional languages and compositionality in agglutinative languages are easy to find. I am confident concluding that the empirical data confirm $\text{Th}_{A'}$ and therefore Th_{A} (i.e., that inflectionality implies idiomatic derivation).

3.2 Th_{B}: idiomaticity ⇒ larger Index∩

Assume $\Lambda=\{L_1, L_2, L_3, ..., L_k\}$ is a set of languages associated to $\mathscr{K}_{\Lambda}$, then:
if [all L_i from Λ have a highly idiomatic derivation],
then [$\mathscr{K}_{\Lambda}$ has a high value of $\mathsf{Index}_{\cap}$].

As already mentioned, the "idiomaticity vs. compositionality" scale is a gradual, rather than binary, feature. Regardless of how the values of this feature are calculated in each specific case, I will take for granted that it is always possible to determine the average value of "idiomaticity/compositionality" of the derivatives in a set of languages participating in certain comparative constructs.

Accordingly, Th_{B} can be reformulated in a somewhat relaxed form:

Thesis B′ ($\text{Th}_{B'}$: idiomaticity ⇒ larger $\mathsf{Index}_{\cap}$). *Other things being equal, the more idiomatic the derivation system of each associate language of a comparative construct, the higher the value of the corresponding* $\mathsf{Index}_{\cap}$.

An analytic justification of $\text{Th}_{B'}$ will be presented in §3.2.1; empirical evidence will be presented in §3.2.2.

3.2.1 Analytic justification of $\mathrm{Th}_{B'}$

The roots located in one row of the cognate table have pairwise similar — or at least correlated — semantics. Only thus can we attribute a particular invariant semantics to the root understood as a row of the cognate table. Recall that, in reality, roots have no inherent semantics even within one language. What we have is just the semantics of the providers (see §1.4).

In the cognate table, columns correspond to associate languages while rows correspond to roots. Each cell of the table contains either the segmental image of the corresponding root in the corresponding language, or is left empty. I wish to distinguish three possible scenarios on such respect.

Scenario 1. The cell is non-empty and semantically consistent, i.e.: the corresponding associate language has a root that is a segmental image of the rest of the roots of the row (pairwise); all, or a large majority, of the lexemes deriving from this root are semantically consistent with the conventional meaning of the row. This is the case with the following row of the IE cognate table:

Skt.	Greek	Lat.	Slav.	Goth.	Arm.	Lith.	TochA.	Hit.
dvā	δύω	*duō*	*dъva*	*twai*	*erkow*	*dù*	*wu*	*dā*

Table 13. Cognates row for PIE **duō*

All these lexemes mean ‘two’, and most of their derivatives also refer to the same semantic field; thus, we have: *dъvakъ* ‘set of two’, *dъvoi* ‘couple’, *dъvoica* ‘pair’ in Old Church Slavonic; *dubius* ‘dual’, *bis* ‘twice’ in Latin; *twisstass* ‘duality’, *tweihns* ‘double’ in Gothic, etc.

Scenario 1 results in a filled cell and increases the value of $\mathsf{Index}_{\cap}$. This scenario is possible both in languages with idiomatic derivation and in those with compositional derivation.

Scenario 2. The cell is empty; i.e., the corresponding associate language presents no root that is a segmental image of the other roots of the row.

This can be observed by the IE root **abel* ‘apple’, which is attested in Celtic, Germanic, Slavic, Baltic groups, but is lacking in Greek, Latin, Hittite, Indo-Iranian, and Tocharian:

Old Irish	English	Lith.	Russian	Latin	Greek	Hittite	Ind.-Ir.	Toch.
ubull	*apple*	*obuolas*	*jabl.ok.o*	—	—	—	—	—

Table 14. Cognates row for PIE **abel*

Scenario 2 obviously results in a decrease of the value of $\mathsf{Index}_{\cap}$. It can equally occur in both type of languages.

Scenario 3. The cell is non-empty but semantically deviant, i.e.: the corresponding associate language presents a root that is a segmental image of all the other roots of the row (pairwise); however, lexemes deriving from this root and being semantically consistent with the row's conventional meaning are rare, if any, in this language.

Consider the following cognate row, showing the segmental image of the root, together with the corresponding provider and its meaning, for each associate language; notice particularly the Latin column:

Sanskrit	Avestan	Greek	Latin	Hittite	German	Welsh
pat	*pat*	πετ	*pet*	*pat*	*fed*	*hed*
pátati 's/he flies'	*pata^{i}ti* 's/he flies'	πέτομαι 'I fly'	[see below]	*pattar* 'wing'	*Feder* 'pen'	*hedant* 'they fly'

Table 15. Cognates row for PIE **pet*/*pt*

Latin root *pet*/*pt* derives a large number of lexemes (see de Vaan 2008: 463–464); below, they are listed according to the three semantic fields they belong to:

(i) *penna* (*pet.na*) 'pen', *praepes* (*prae.pet.s*) 'flying forward';

(ii) *pet.ere* 'ask', *pet.itio* 'petition', *pet.ulans* 'pert, petulant', etc.;

(iii) *pro.pt.ervus* 'bold, violent', *im.pet.us* 'rapid motion, attack'.

It is clear that only group (i) shows some semantic similarity with the rest of the provider-lexemes of the IE cognate row shown in Table 15; groups (ii) and (iii) are, apparently, totally deviant from the row's conventional meaning.

Scenario 3 is trickier comparing to the first two. It appears highly unlikely, or even impossible, in languages with compositional derivation, whereas, on the other hand, it is quite normal and widely attested in languages with idiomatic derivation.

Consider a group of same-root derivatives in a language with compositional derivation. All such lexemes would be semantically consistent. The compositionality prevents their meanings from diverging from one another. Consequently, either all such lexemes could work as root providers for a cell in a cognate table, or none of them could.

In languages with idiomatic derivation, on the other hand, the semantic divergence of some lexemes in a group of same-root derivatives is fairly possible. The fact that some words have departed from the original semantics of the group does not imply that the group as a whole must change its semantics. Therefore, there would still be some lexemes that preserve the original semantics, available as providers for a cell in a cognate table, given the expected row's "meaning". This is exactly the case with the Latin example mentioned above (p. 25): although the words of the subgroups (ii) and (iii) are no longer semantically consistent with the row's conventional "meaning", there are still the words of the subgroup (i) — such as *penna* (from *pet.na*) — that are fully consistent with the desired semantics of "pen/wing/fly".

In both idiomatic and compositional languages the words are, so to speak, "washed out" and disappear from the lexicon altogether over time. This process is totally physiological, and eventually creates empty cells in the cognate table. Yet, the mechanism of this "washout" is different in the two types of languages. In languages with idiomatic derivation, each lexeme in a group of same-root derivatives has, so to speak, its own "destiny": it changes its meaning or simply disappears from the lexicon independently of the rest of the group. This is particularly evident when it is the basic lexeme that disappears, leaving ***orphaned*** derivatives. For example, English has *be.gin* but no **gin*, *hund.red* but no **hund* (cf. *hat.red* from *hate*); Spanish has *pro.ducir*, *a.ducir*, *con.ducir*, etc, but no **ducir*.

Conversely, in languages with compositional derivation, lexemes with a common root have, so to speak, a "common fate": they change their meanings, or disappear, as one whole cohort, leaving no orphans. This is why Scenario 3 is simply impossible in languages with compositional derivation: such languages lose their roots abruptly. Where there may only be the loss of one word in languages with idiomatic derivation, in compositional languages there will be a loss of the entire group of same-root derivatives; i.e., the loss of the root.

3.2.2 Evidence

Consider again two of the propositions discussed above, H_2 and $Th_{B'}$:

H_2 The Indo-European comparative construct exhibits a high $\mathsf{Index}_{\cap}$ given the inflectionality of its associate languages.

$Th_{B'}$ Other conditions being equal, the more idiomatic the derivation system of each associate language of a comparative construct, the larger the value of the corresponding $\mathsf{Index}_{\cap}$.

All inflectional languages under consideration in the present Section have a highly idiomatic derivation, while all agglutinative languages have a highly compositional derivation (see Th_A, already demonstrated analytically and empirically substantiated in §3.1). Therefore, the empirical validation of both $Th_{B'}$ and H_2 (which, in turn, implies the truth of H_1) consists in providing the same evidence.

Let us analyse such evidence. For instance, I must show that $\mathsf{Index}_{\cap}$ of agglutinative languages is lower than $\mathsf{Index}_{\cap}$ of inflectional languages. To comply with the "other things being equal" condition, I will consider equal numbers of associate languages, with as similar values of $\mathsf{Index}_{\cup}$ and ranges of $\mathsf{Index}_{\updownarrow}$ as possible.

Agglutinative languages

The Uralic family has been chosen to represent the agglutinative type. The Uralic section of *Starling* includes data from 15 associate languages having $\mathsf{Index}_{\cup} = 1898$. In Table 16 they are listed in descending order by $\mathsf{Index}_{\updownarrow}$.

Rank	Language	$\mathsf{Index}_{\updownarrow}$
1	Finnish	927
2	Komi	831
3	Khanty	803
4	Saam	764
5	Estonian	750
6	Mansi	742
7	Hungarian	721
8	Udmurt	718
9	Mordovian	663
10	Mari	647
11	Nenets	405
12	Selkup	357
13	Enets	262
14	Kamass	259
15	Nganasan	247

Table 16. Finno-Ugric languages listed by $\mathsf{Index}_{\updownarrow}$

For my test I will use the five languages with the highest $\mathsf{Index}_{\updownarrow}$ from this group, namely: {Finnish, Komi, Khanty, Saam, and Estonian}. Note that, in fact, these five languages belong to the Finno-Ugric group within the Uralic family, that is, to a compact group of closely related languages with a relatively long-lasting written tradition, good grammatical descriptions and etymological dictionaries.

According to *Starling*, these five languages form a set with $\mathsf{Index}_{\cap} = 109$. The three of them with the highest $\mathsf{Index}_{\updownarrow}$ (namely, {Finnish, Komi, Khanty}) have $\mathsf{Index}_{\cap} = 178$. Again, the three Finno-Ugric languages with the most developed written tradition and, therefore, the best elaborated grammatical descriptions and etymology, namely {Finnish, Hungarian, and Estonian}, with ranks (1, 5, 7), have the value of $\mathsf{Index}_{\cap} = 245$.

Inflexional languages

Germanic languages have been chosen as representatives of the inflectional typology. They form a compact group of closely related languages within the IE family, not unlike the Finno-Ugric languages within the Uralic family. Data for 21 Germanic languages are included in a special section of *Starling*. They present $\mathsf{Index}_{\cup} = 1991$. On Table 17 (p. 28) they are listed in descending order by $\mathsf{Index}_{\updownarrow}$.

For our purpose it would be inappropriate to take the "best" five Germanic languages, since their $\mathsf{Index}_{\updownarrow}$ values are all significantly higher than those of the "best" five Finno-Ugric languages. Therefore, we must take a subset of Germanic that better com-

Rank	Language	$\mathsf{Index}_{\updownarrow}$
1	Old English	1430
2	Middle High German	1376
3	Old Norse	1372
4	Old High German	1363
5	German	1251
6	Norwegian	1166
7	Middle Low German	1163
8	Dutch	1161
9	Middle Dutch	1144
10	Swedish	1089
11	Old Saxon	941
12	English	912
13	Danish	856
14	Gothic	689
15	Old Frisian	666
16	Old Swedish	356
17	Old Frankish	263
18	Old Danish	248
19	Low Geman	166
20	East Frisian	119
21	Middle English	117

Table 17. Germanic languages listed by $\mathsf{Index}_{\updownarrow}$

plies with the parity requirement, which is: {Old Saxon, English, Danish, Gothic, Old Frisian}, i.e. the interval 11–15. This subset has the value of $\mathsf{Index}_{\cap} = 235$; the first three of them (11–13) have the value of $\mathsf{Index}_{\cap} = 440$.

Table 18 sums up the results of these simple calculations. The rows correspond to the three measurements that have been made. In the first measurement two subsets of the "best" five languages are compared. The second one compares the subsets of the three "best" languages from each group. The last one compares the subset of the three "best" Germanic languages with the three literary Finno-Ugric languages.

No.	Germanic		Uralic	
	Sets	$\mathsf{Index}_{\cap}$	Sets	$\mathsf{Index}_{\cap}$
1	11–15	235	1–5	109
2	11–13	440	1–3	178
3	11–13	440	1, 5, 7	245

Table 18. Measurements

As we can see, whatever subset of Germanic languages we consider, it will always show a value of $\mathsf{Index}_{\cap}$ that is around twice as much as the corresponding value for a similar subset Uralic languages.

Therefore, the evidence allows us to conclude that the initial hypothesis is confirmed by the data.

4 Conclusions

Summing up the argumentation of the preceding Sections, I conclude that, provided that the reliability of a comparative construct is interpreted as the number of common roots, this number is correlated with the degree of idiomaticity of derivation in the associate languages, which, in turn, is correlated with the degree of their inflectionality (in a special sense of this term). The test performed on a group of closely related Indo-European languages (i.e., Germanic) and a group of closely related Uralic languages (i.e., Finno-Ugric) shows that the data confirm the hypothesis. In selecting the data to analyse, my main concern was to obtain equal starting conditions (number of languages, "richness" of languages) in two random groups of genetically related languages. Now, the fact that *Starling* has a special section with Germanic data is a totally random outcome of how this database had been developed. The fact that the Finno-Ugric group of languages, as they are accounted on *Starling*, present starting conditions similar to those of the Germanic ones is, again, a totally random condition. Therefore, I feel safe to conclude that my choice of samples has the required degree of randomisation, for a meaningful comparison.

Therefore, the inflectionality may, at least partly, explain the major empirical reliability of the Indo-European comparative construct and its role in historical linguistics. As a matter of fact, most of the other linguistic families for which comparative constructs have been ever formulated happen to be either agglutinative (in the sense defined in the present paper), or isolating, in which case the methodology presented here would yield even lower figures.

It is also possible that in my analysis of the data from Germanic and Finno-Ugric languages I have overlooked some less obvious marginal factors which might have influenced the results. Thus, it can easily occur that the date of the genetic split between the languages under observation plays a role in determining the $\mathsf{Index}_{\cap}$ of these languages. Thus, closely related languages (as is the case with the Germanic ones) could anyway show a higher value of $\mathsf{Index}_{\cap}$ in comparison to distantly related languages (such as the Ugro-Finnic ones), regardless their grammatical typology. Further analysis is needed in order to address this issue. But, whatever the final outcome of this future investigation might be, the fact remains that the value of $\mathsf{Index}_{\cap}$, as defined in the present paper, remains a valid metrics for a meaningful comparison between different genetic hypothesis in historical linguistics.

Acknowledgements

My kind thanks to P. Arkadiev, L. Blumenfeld, C. Ciancaglini, C. D'Antonio, A. del Tomba, B. Kravcov, and A. Polivanova for reading a preliminary version of this paper: their useful remarks helped me avoiding some unfortunate errors and inaccuracies.

References

Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511486210.

Bertinetto, Pier-Marco. 1995. Compositionality and Non-compositionality in Morphology. In W. Dressler & C. Burani (eds.), *Crosslinguistic Approaches to Morphology*, 9–36. Vienna: Verlag der Osterreichischen Akademie der Wissenschaften.

Brečalova, E. V. 2009. *Principy postroenija sintaktičeskogo predstavlenija korejskogo predloženija*. Moscow: Institut Jazykoznanija RAN dissertation.

Bucsko, John Martin. 2011. *Preverbs and idiomatization in Gothic*. New York: Lang.

Chomsky, Noam A. & Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.

Ciancaglini, Claudia A. 2009. How to prove genetic relationships among languages: the cases of Japanese and Korean. *Rivista degli Studi Orientali* 81. 289–320.

Ciancaglini, Claudia A. 2012. Il suffisso indo-ir. *-*ka*- nelle lingue iraniche antiche. *Archivio Glottologico Italiano* 98(1). 3–33.

Cruse, D. Alan. 2000. *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Dekker, Peter & Willem Zuidema. 2020. Word prediction in computational historical linguistics. *Journal of Language Modelling* 8(2). 295–336. DOI: 10.15398/jlm.v8i2.268.

Doerfer, Gerhard. 1981. The conditions for proving the genetic relationship of languages. *The Bulletin of the International Institute for Linguistic Sciences, Kyoto Sangyo University* 2(4). 39–58.

Dyen, Isidore. 1969. Reconstruction, the Comparative Method, and the Proto-Language Uniformity Assumption. *Language* 45(3). 499–518.

Ellis, Jeffrey. 1953. *An Elementary Old High German Grammar*. Oxford: Clarendon Press.

Frellesvig, Bjarke. 2010. *A History of the Japanese Language*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511778322.

Hakulinen, Lauri. 1961. *The Structure and Development of the Finnish Language*. Trans. from the Finnish by John Atkinson. The Hague: Mouton.

Haspelmath, Martin. 2009. An Empirical Test of the Agglutination Hypothesis. In Sergio Scalise et al. (eds.), *Universals of Language Today*, 13–29. Rotterdam: Springer. DOI: 10.1007/978-1-4020-8825-4_2.

Hoeksema, Jack. 2000. Compositionality of meaning. In Geert Booij, Christian Lehmann & Joachim Mugdan (eds.), *Morphology. A Handbook on Inflection and Word-Formation*, vol. 1, 851–857. Berlin & New York: Walter de Gruyter. DOI: 10.1515/9783110111286.1.11.851.

Hoenigswald, Henry M. 1973. *Studies in Formal Historical Linguistics*. Dordrecht: D. Reidel.

Holthausen, Ferdinand. 1934. *Gotisches etymologisches Wörterbuch: mit Einschluss der Eigennamen und der gotischen Lehnwörter im Romanischen*. Heidelberg: Winter.

Karlsson, Fred. 2018. *Finnish. A Comprehensive Grammar*. London & New York: Routledge. DOI: 10.4324/9781315743547.

Krahe, Hans. 1960. *Germanische Sprachwissenschaft: I. Einleitung und Lautlehre*. Berlin: De Gruyter.

Kroonen, Guus. 2013. *Etymological Dictionary of Proto-Germanic*. Leiden & Boston: Brill.

Partee, Barbara H. 1984. Compositionality. In F. Landman & F. Veldman (eds.), *Varieties of Formal Semantics*, 281–311. Dordrecht: Foris.

Plungjan, Vladimir. 2001. Agglutination and flection. In M. Haspelmath et al. (eds.), *Language typology and language universals. An international handbook*, vol. 1, 669–678. Berlin: Mouton de Gruyter.

Pokorny, Julius. 1959. *Indogermanisches etymologisches Wörterbuch*. Bern: Francke.

Polivanova, A. K. 2008. *Obščee i russkoe jazykoznanie. Izbrannye raboty*. Moskva: Rossijskij Gosudarstvennyj Gumanitarnyj Universitet.

Polivanova, A. K. 2013. *Staroslavjanskij jazyk. Grammatika. Slovari*. Moskva: Universitet Dmitrija Požarskogo.

Polivanova, A. K. 2022. *Logičeskie osnovy grammatiki. Ot fonologii do semantiki*. Moskva: Izdatel'skij dom Vysšej Školy Ėkonomiki.

Rounds, Carol. 2001. *Hungarian. An Essential Grammar*. London & New York: Routledge.

Singh, Arindama. 2009. *Elements of Computation Theory*. Dordrecht: Springer. DOI: 10.1007/978-1-84882-497-3.

Starostin, Sergei et al. 1998–2013. *Tower of babel. Etymological databases*. https://starling.rinet.ru.

de Vaan, Michiel. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Leiden & Boston: Brill.

Vasmer, Max Julius Friedrich. 1950–1958. *Russisches etymologisches Wörterbuch*. Heidelberg: C. Winter.

Vovin, Alexander. 2005. The end of the Altaic controversy. *Central Asiatic Journal volume* 49(1). 71–132.