Binominal fingerprints and Indo-European genetics

Steve Pepper
University of Oslo

Keywords: typology; language comparison; nominal modification; binominal lexemes; compounding

This paper examines the potential of binominal lexemes in computer-assisted approaches to historical and typological language comparison. It describes an onomasiologically structured lexical database and the use of statistical clustering to reveal patterns of genetic affiliation and language contact.

The comparative concept encompasses noun-noun compounds and their functional equivalents. A binominal lexeme is a “lexical unit that consists primarily of two thing-morphs and whose function is to name a complex concept that involves an unstated (or under-specified) relation between two entities” (Pepper to appear). Indo-European examples denoting RAILWAY are given in (1).

(1) (a) German Eisen-bahn (b) French chemin de fer (c) Russian ⟭ bileznaja doroga ⟭ żelez-n-ica (d) Slovak iron-way way of iron iron-ADJZ way iron-ADJZ-NMLZ

The binominals database is built around a set of 100 concepts, 84 of which are found in the Concepticon (List et al. 2018), and their denotations in 106 languages. The data were extracted from WOLD (Haspelmath & Tadmor 2009) and supplemented using dictionaries and questionnaires. Each binominal is annotated for construction (e.g. Ger. Mod Head, Fr. Head PREP Mod, etc.); order of constituents (here, R and L, respectively); semantic relation (here, COMPOSITION); and morphosyntactic strategy. The latter uses an eight-way typology: four types are exemplified in (Error! Reference source not found.): (a) compounding (cmp), (b) prepositional (prp), (c) adjectival (adj) and (d) derivational (der); the remaining four strategies are exemplified in (2) by items denoting NOSTRIL: (a) genitival (gen), (b) construct (con), (c) double-marking (dbl) and (d) classifier (cls).

(2) (a) Archi muč-li-n klan (b) Turkish burun deliğ-i (c) Takia ṱdu-n awa-n (d) Murui Huitoto defo
  nose-OBL-GEN hole nose hole-3SG nose-3SG mouth-3SG nose.CL(cavity)

Statistics show that languages vary greatly in terms of the number and kinds of morphosyntactic strategies used: from one in Norwegian (cmp) to as many as five in Polish (cmp, prp, gen, adj, and der). Where there is competition, as in Polish, the relative frequency also varies. The data are exploited to create a binominal profile or “fingerprint” for each language, notated as [100cmp] for Norwegian, [4cmp7prp5gen34adj50der] for Polish, etc. These turn out to be highly indicative of genetic affiliation, at least in Indo-European, the family for which most data is currently available (Figure 1). Germanic, Romance and Slavic in particular show clear intragroup similarities and equally clear intergroup differences.

---

1 A thing-morph is a root or affix that profiles a thing, cf. Haspelmath’s (2012) notion of ‘thing-root’.
This is confirmed by statistical clustering performed using Principal Components Analysis (Figure 2). Where languages from the same genus do not cluster together, as is the case with Celtic and Indo-Aryan languages, explanations are sought in grammaticalization and language contact, as will be shown in the talk. Thus the formal patterning of the binominal lexicon, arrived at by computer-assisted means, tells a story about the linguistic history of the language in question.

References

Hapelmeh, Martin. 2012. How to compare major word-classes across the world’s languages. UCLA Working Papers in Linguistics, Theories of Everything 17, Article 16. 109–130.


---

2 In the diagram, each of the six quadrants containing symbols provides a view into the same three-dimensional space of binominal types, as though looking through the different faces of a cube.