Formal Concept Analysis meets grammar typology

Kow Kuroda
kow.k@ks.kyorin-u.ac.jp
Medical Department, Kyorin University

1 Introduction

How to work out a **typology of grammar** that can help language learners to figure out the grammar features/attribute that they need most seriously at learning a (new) language? If such a typology exists, it surely will be helpful but it requires measuring the similarity among grammars of language rather than their vocabularies. How to implement the idea? Everybody knows that grammar is a very complex entity and it's really hard to rank features relevant to it.

So-called "language distance" turns out not to be as useful as expected, because it usually measures the rate of shared vocabulary elements. This could put phylogenic bias into grammar classification, so that languages with different types tend to be classified as close when they are branching off the same root. Discussions of grammar typology tend to rely on diachronic evidence. This is simply because we are convinced, perhaps correctly, that grammars have evolved.

Does this mean that it is impossible to think of grammar typologies solely based on synchronic evidence and to expect them to be useful in actual learning process of a language? This is the very question we want to address in this paper. And the suggested answer is positive.

2 Data and Analysis

The 15 languages listed in (1) were selected:

(1) Bulgarian, Chinese, Czech, English, French, Finnish, German, Hebrew, Hungarian, Japanese, Korean, Latin, Russian, Swahili, and Tagalog

After dozens of tests with trial and error, the following 24 attributes were selected:[1)]

(2) **A1** HAS DEFINITE ARTICLE; **A2** HAS INDEFINITE ARTICLE; **A3** N ENCODES PLURALITY; **A4** N ENCODES CLASS; **A5** N ENCODES CASE; **A6** RELATIVE CLAUSE FOLLOWS N; **A7** HAS POSTPOSITIONS; **A8** HAS PREPOSITIONS; **A9** A AGREES WITH N-PLURALITY; **A10** A AGREES WITH N-CLASS; **A11** A AGREES WITH N-CASE; **A12** A FOLLOWS N; **A13** O MUST FOLLOW V; **A14** REQUIRES SUBJECT; **A15** V ENCODES VOICE; **A16** V ENCODES TENSE; **A17** V ENCODES ASPECT; **A18** V AGREES WITH SUBJECT; **A19** V ENCODES PERSON; **A20** V ENCODES PLURALITY; **A21** V ENCODES N-CLASS; **A22** V INFINITIVE IS DERIVED; **A23** V AGREES WITH OBJECT; **A24** HAS TENSE AGREEMENT

Admittedly, the selection of languages in (1) is biased for well-documented languages with attributes that amend manually checking, but its was done with two goals in mind. First, it aims to cover as wide a variety of languages as possible. Second, it aims to include as many phylogenically unrelated languages as possible. The second point is important because this research looks for patterns that would suggest convergent evolution of grammars without reference to the evolution of human languages that has actually occurred.

Formal Concept Analysis (FCA) (Ganter and Wille 1999) was used for analysis. FCA is a mathematical approach to classification and is proved to be powerful enough to handle data with essential complexity efficiently. Building on lattice theory, it works on a "formal context" and produces a "Hasse diagram" as result.

The formal context in Figure 1 was manually prepared and FCA was applied to it.[2)] `ConceptExplorer 1.3`[3)] was used to perform FCA with drawing options: `Layout = minimal intersection`, `Draw mode = ∼stability` (other options don't affect results).

3 Results

We compare two kinds of result. One kind is FCA in which all attributes are used, and therefore the result is "uncompromised." Another kind is FCA in which a selection of attributes are used, and therefore the result is "compromised." This selection is done, through trial and error, to optimize of the results. `ConceptExplorer 1.3` has an option `Show collisions` to assist this process.

Figure 2 shows the "uncompromised" FCA. Compared to "compromised" FCAs, this form of FCA is not really revealing, but some note would be helpful about it.

First, the lower in the Hasse diagram a language is, the more feature-loaded its grammar is. In this respect, languages directly connected to the bottom such as Swahili, Hebrew, Hungarian, Russian/Czech, Finnish, French, Latin, German, are languages with most "feature-loaded" (and complex) grammars. On the other hand, Chinese has a less feature-loaded (and simple) grammar.[4)]

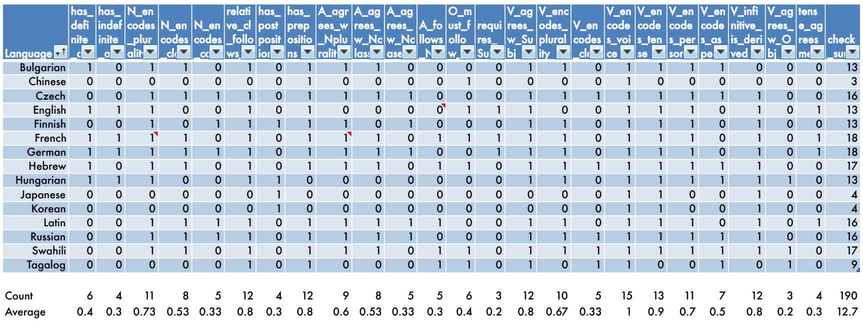
Second, uncompromised use of all features rarely yields a good analysis. The following criteria are presumed to assess

[1)] After experiments, a decision was made to let (N) CLASS to include GENDER, knowing that this may deviate from usual practice in linguistics. This means that GENDER is treated in the same way as "Noun classes" found in languages like Tagalog and Swahili.

[2)] Many details of the attribute-value pairs are admittedly debatable and are open to critical assessment in that some of them look illegitimate to salted linguists.

[3)] `http://conexp.sourceforge.net/`

[4)] It is out of the scope of this research to ask what determines the complexity of a grammar, but it is worth a mention that the size of speaker population correlates with it (Lupyan and Dale 2010).



Language	has_ defi nite _d	has_ indef inite _a	N_en codes _plur alit	N_en codes _cl	N_en codes _cc	relati ve_cl _follo ws	has_ post posit ion	has_ prep ositio ns	A_agr ees_w _Nplu ralit	A_ag rees_ w_Nc lass	A_ag rees_ w_Nc ase	A_fo llows _N	O_m ust_f ollo w_	requi res_ Su	V_ag rees_ w_Su bj	V_enc odes_ plural ity	V_en codes _cl	V_en code s_voi ce	V_en code s_ten se	V_en code s_per sor	V_en code s_as pe	V_infi nitive_ is_deri ved	V_ag rees_ w_O bj	tens e_ag rees me	check _su
Bulgarian	1	0	1	1	0	1	0	1	1	1	0	0	0	0	1	1	0	1	1	1	1	0	0	0	13
Chinese	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	3
Czech	0	0	1	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	0	0	16
English	1	1	1	0	0	1	0	1	0	0	0	0	1	1	1	0	0	1	1	1	0	1	0	1	13
Finnish	0	0	1	0	1	1	1	1	1	0	1	0	0	0	1	1	0	1	1	1	0	1	0	0	13
French	1	1	1	1	0	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	0	1	0	1	18
German	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1	1	0	1	1	1	0	1	0	1	18
Hebrew	1	0	1	1	0	1	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	0	0	17
Hungarian	1	1	1	0	0	1	1	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	0	13
Japanese	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	4
Korean	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	4
Latin	0	0	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	1	1	1	0	1	0	1	16
Russian	0	0	1	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	0	0	16
Swahili	0	0	1	1	0	1	0	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	0	17
Tagalog	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	1	1	0	0	1	0	1	0	9
Count	6	4	11	8	5	12	4	12	9	8	5	5	6	3	12	10	5	15	13	11	7	12	3	4	190
Average	0.4	0.3	0.73	0.53	0.33	0.8	0.3	0.8	0.6	0.53	0.33	0.3	0.4	0.2	0.8	0.67	0.33	1	0.9	0.7	0.5	0.8	0.2	0.3	12.7

Figure 1: **formal context encoding 15 languages with 25 attributes**

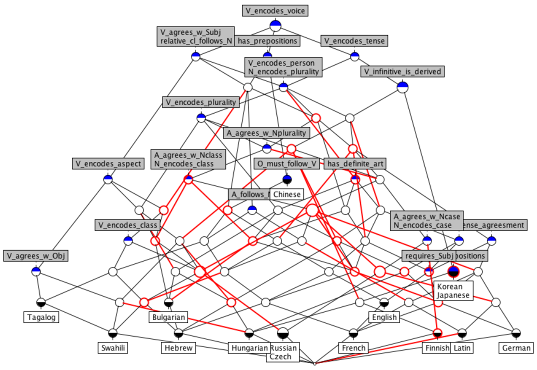


Figure 2: **Uncompromised FCA**: red indicates "collision"

the goodness of FCA: A Hesse diagram is good if i) objects are as much separated as possible (**condition 1**), but ii) there are as few empty nodes as possible (**condition 2**), and iii) the diagram is in a geometrically good shape (**condition 3**). For example, the Hasse digram in Figure 2 fails to meet conditions 2 and 3.

The three conditions cancel each other, because they are in relation of "trade-off" that need to be "compromised." This means that the there is no fast and sure way to obtain the optimal output by FCA. Thus, we need to manually compare several results generated under different selections of attributes to obtain optimal ones. This is what is attempted in the following.

3.1 Optimization 1

Optimization can be achieved by discarding either objects or attributes. In our analysis, all objects are expected to be sufficiently reliable and were retained.

The Hasse diagram in Figure 3 is one of the "compromised" FCAs. We contend that this is the best optimization in that conditions 1 and 3 are fully satisfied at cost of discarding condition 2, as far as we see reason that accidental gaps in data, either in terms of object or attribute, often result in empty nodes.

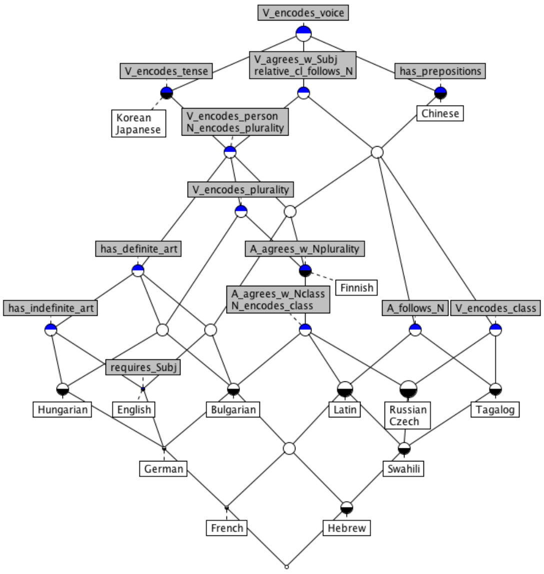


Figure 3: **Compromised FCA 1 with 5 empty nodes**

3.1.1 Feature interactions

Figure 3 presents the result of separating attributes into "converging" attributes in (3) and "diverging" attributes in (4):

(3) **A1** HAS DEFINITE ARTICLE; **A2** HAS INDEFINITE ARTICLE; **A3** N ENCODES PLURALITY; **A4** N ENCODES CLASS; **A6** RELATIVE CLAUSE FOLLOWS N; **A8** HAS PREPOSITIONS; **A9** A AGREES WITH N-PLURALITY; **A10** A AGREES WITH N-CLASS; **A12** A FOLLOWS N; **A14** REQUIRES SUBJECT; **A15** V ENCODES VOICE; **A16** V ENCODES TENSE; **A18** V AGREES WITH SUBJECT; **A19** V ENCODES PERSON; **A20** V ENCODES PLURALITY; **A21** V ENCODES N-CLASS;

(4) **A5** N ENCODES CASE; **A7** HAS POSTPOSITIONS; **A13** O MUST FOLLOW V; **A17** V ENCODES ASPECT; **A22** V INFINITIVE IS DERIVED; **A23** V AGREES WITH OBJECT; **A24** HAS TENSE AGREEMENT;

Attributes in (4) are called "divergent," because inclusion of them inevitably adds undesirable complications to the output. The best account for this would be that certain attributes get inconsistent values and thus contradictions are generated. But this is not necessarily due to "errors" in encoding. §4.3 discusses it more deeply.

3.1.2 Implications among attributes

The FCA in Figure 3 expresses the following:

(5) Correlations (stated bottom-up)

a. Two attributes, **A4** N encodes class and **A10** A agrees with N-class, correlate, if not equivalent.

b. Two attributes, **A19** V encodes person, and **A20** V encodes plurality, correlate, if not equivalent.

c. Two attributes **A6** relative clause follows N, and **A18** V agrees with Subject, correlate, if not equivalent.

(6) Implications (stated bottom-up, indirect implications not expanded)

a. **A2** has indefinite article is a precondition for **A14** requires Subject.

b. **A1** has definite article is a precondition for **A2** has indefinite article.

c. **A9** A agrees with N-plurality is a precondition for **A4** N encodes class and **A10** A agrees with N-class.

d. **A20** V encodes plurality is a precondition for **A4** N encodes class, **A9** A agrees with N-plurality, and **A10** A agrees with N-class.

e. **A19** V encodes person and **A3** N encodes plurality are a precondition for **A20** V encodes plurality.

f. **A8** has prepositions is a precondition for **A14** requires Subject, **A9** A agrees with N-plurality, **A12** A follows N, and **A21** V encodes N-class.

g. **A15** V encodes voice and **A6** relative clause follows N are a precondition for **A16** V encodes tense, **A3** N encodes plurality, **A12** A follows N, and **A18** V agrees with Subject.

h. **A16** V encodes tense is a precondition for **A19** V encodes person and **A3** N encodes plurality.

3.1.3 Beyond Greenberg's universals

Obviously, (5) and (6) re-capture some of **Greenberg's language universals** (Greenberg 1966), but the Hasse diagram in Figure 3 tells more, provided that it is a correct analysis. What FCA gives us is not a "list" of implications, but a complex "space" encoding how attributes are combined to **define grammar types** in a hierarchical fashion, thereby providing something like a "geometry" of grammars.

3.2 Comparison with other optimizations

In what follows, we compare the FCA in Figure 3 against other possibilities to defend that it is the best result under the available data. Note, however, that the comparison is not intended to be exhaustive. Space limitation discourages us to investigate all the relevant combinations of attributes even if it is practically possible.

3.2.1 Optimization 2

Figure 4 presents Optimization 2 in which **A20** V encodes plurality is discarded additionally, generating 4 empty nodes. This Hasse diagram involves four empty nodes, without overloaded nodes. The major difference from Figure 3 is that the geometry is simpler and less symmetrical, suggesting the result is sub-optimal.

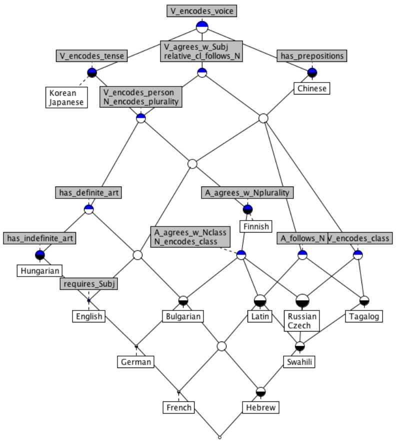


Figure 4: **Compromised FCA 2 with 4 empty nodes**

3.2.2 Optimization 3

Figure 5 presents Optimization 3 in which **A1** and **A9** are discarded additionally, generating 3 empty nodes. This Hasse diagram involves three empty nodes, without node-overloading. But the geometry is less symmetrical.

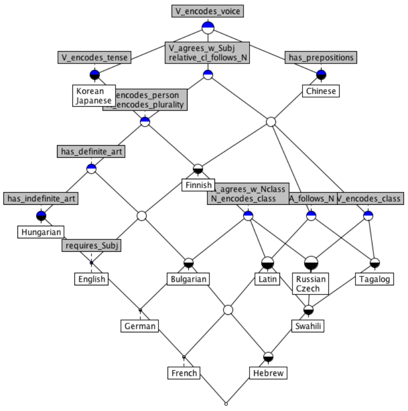


Figure 5: **Compromised FCA 3 with 3 empty nodes**

3.2.3 Optimization 4

Figure 6 presents Optimization 4 in which **A9**, **A12** and **A20** are discarded additionally, generating 2 empty nodes. This Hasse diagram involves two empty nodes, with overloaded nodes: at one node, Swahili and Russian/Czech are undifferentiated; at another, French and German undifferentiated. This suggests that the result is not optimal.

3.2.4 Optimization 5

Figure 7 presents Optimization 5 without empty nodes under the selection of attributes in **A3**, **A4**, **A5**, **A6**, **A7**, **A8**, **A9**, **A10**, **A11**, **A15**, **A16**, **A18**, **A19**, and **A20**.

This Hasse diagram is free from empty nodes, but some nodes are overloaded. This means that certain objects are under-represented and FCA fails to differentiate them sufficiently. This means that this result is not optimal.

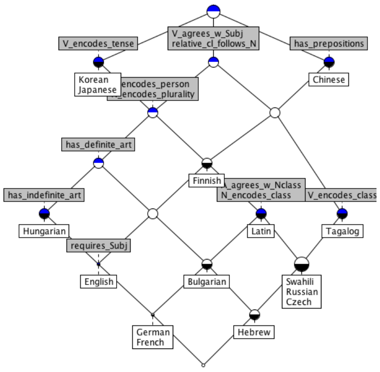


Figure 6: **Compromised FCA 4 with 2 empty nodes**

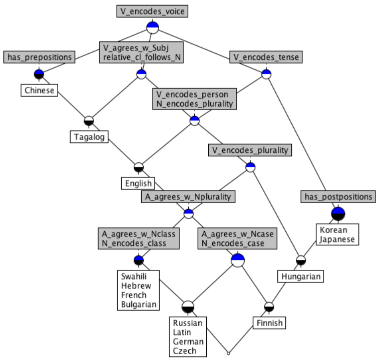


Figure 7: **Compromised FCA 5 without empty nodes**

4 Discussion

4.1 Relativized learnability index

The Hasse digram in Figure 3 presents something like a "common ground" of grammars: **A18** V agrees with Subject (and **A19** V encodes person as one of its most accessible incarnation). This feature is so important that we can suspect a large and deep gap between inside and outside the domain in which it holds. Only Chinese, Japanese and Korean are outside the domain. Another influential attribute is **A3** N encodes plurality. Many languages have grammars sensitive to it. This would create another barrier in language learning.

This has important implications for language learning. In fact, it is imaginable that learners face more difficulty if their mother tongue is one of the agreement-free languages. If a learner speaks a language without person-agreement on verbs and plurality-encoding on nouns, it would pose a handicap in his or her learning. We can reasonably predict that, other things being equal, descending the Hasse diagram poses more difficulty in learning. This defines **relativized learnability index** for grammar.

4.2 Caveat on the nature of representation

FCA is a powerful and useful tool to reveal about grammar types; yet proper interpretation of its results demands additional explanation. In the Hasse diagrams, grammar types are represented as discrete objects. We are discourage to understand the representation at its face value. Most notably, encoding of attributes suffers from abstractions at several levels. For one, grammatical categories like Noun, Verb Adjective are abstractions. In reality, each of them subsumes a group of words that behave differently. For another, V subsumes different word classes such as Present, Past, Participle, Perfective, Imperfective, Infinitive. The same is very true of Adjectives, too. For yet another, the operational definition Case is problematic, to say the least. Also, it is not clear how far the notion Noun class should cover. All this encourages us to reinterpret what the Hasse diagrams represent more probabilistically. Perhaps, grammar types are best understood as "attractors" in a dynamical system, in analogy with "niches" over a "fitness" landscape.

4.3 Why divergent attributes?

Why are some attributes divergent? We can see two possibilities for this, on different grounds. First, it is rather likely that certain attributes were specified for "wrong" values in the formal context used. This would have produced inconsistencies in encoding. Why is this the case? The answer would be, at least in part, that certain grammatical phenomena are ill-defined. Case, for example, turned out to be too unreliable an attribute to add complications, thereby suggesting that its identification involves essential difficulty.

There is another possibility, however, which is of more theoretical importance. For any existing language, there is a chance that attested grammatical features are accidentally valued, or even spurious, and there is no systematic way to decide which ones are the case. Careful choice of relevant features is essential for valid generalizations.

After all, language would be a "system of trade-offs" that involves counterbalancing a large number of costs and benefits. Why is grammar not so? If this line of thought is valid, it will be completely misguided to try to think of grammar as a "systematic" phenomenon and expect it to be properly characterizable by a "monolithic" system of rigid "rules and principles." Knowledge of language must be much more probabilistic, if not stochastic, and involves much more complexity in it.

5 Conclusion

This paper presented a FCA-based analysis of grammar typology in terms of attributes/features such as **A19** V encodes person, The presented approach gives promising results that capture implications among attributes of grammar and automatically identifies grammar types, though the results are preliminary in many respects.

References

Ganter, B. and R. Wille (1999). *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer-Verlag.

Greenberg, J. H. (1966). *Language Universals: With Special Reference to Feature Hierarchies*. The Hague: Mouton.

Lupyan, G. and R. Dale (2010). Language structure is partly determined by social structure. *PLoS ONE 5*(1), 1–10.