# How are inflectional paradigms represented (in the mind)?

## Formal Concept Analysis meets Czech declensional paradigms

Kow Kuroda*
School of Medicine, Kyorin University

### 1 Introduction

There are many languages in the world in which nouns (and possibly adjectives) inflect to CASES like Accusative, Genitive, Dative, etc.. This phenomenon is called "declension" on nouns (and adjectives). If words *X* and *Y* do not share a declensional pattern, they are said to follow different "(declensional) paradigms."[1] Czech is one of such languages, in which, for example, *okno* (window) and *místo* (city, town) have different paradigms. The difference between their declensions in Tables 1 and 2 should reveal this. They have different endings, *(okn)-**u*** and *(míst)-**o*** for Dative singular, though the other cases have the same endings.[2]

Table 1: Declension(s) of *okno* (window)

| CASE | Nom. | Gen. | Dat. | Accu. | Voc. | Loc. | Inst. |
|---|---|---|---|---|---|---|---|
| SING. | *okno* | *okna* | ***oknu*** | *okno* | *okno* | *okně* | *oknem* |
| PLUR. | *okna* | *oken* | *oknům* | *okna* | *okna* | *oknech* | *okny* |

Table 2: Declension(s) of *místo* (city, town)

| CASE | Nom. | Gen. | Dat. | Accu. | Voc. | Loc. | Inst. |
|---|---|---|---|---|---|---|---|
| SING. | *místo* | *místa* | ***místu*** | *místo* | *místo* | *místě* | *místem* |
| PLUR. | *místa* | *míst* | *místům* | *místa* | *místa* | *místech* | *místy* |

Many other languages exhibit declension: Russian does, Latin does, and so on. So, the phenomenon is not rare, if not popular. But it has been long unknown how to represent/encode a particular declensional paradigm formally, keeping multiple contrasts among a good number of different classes of declension. This is because the recognition of a paradigm, or inflectional class in general, cannot be equated with the recognition of particular word forms. Declension describes a second-order entity, the way surface word forms are related to each other.

Here are crucial questions addressed in this paper: **What kind of information is necessary and sufficient to identify particular paradigms as such?** And also, **How is it possible for a speaker to identify particular declensional paradigms as such?** This paper aims to answer these questions, relying on Formal Concept Analysis (FCA) (Ganter and Wille 1999). But the results to be obtained are not limited to declensions.

### 2 Data and Analysis

#### 2.1 Data

Czech nouns and adjectives have dozens of declensional classes, which are simply referred to as "declensions" hereafter. The exact number of them depends on a theory that one assumes, but *English-Czech & Czech-English Compact Dictionary* (Fronek 2010), for example, lists 57 classes for common nouns, or 78 classes if pronouns and adjectives are included. These classes were encoded in the way specified below, and then FCA was applied to it. A reference book (Kanazashi 1998) was also consulted to construct the database used. To complete missing information, `https://en.wikipedia.org/wiki/Czech_declension` was consulted as a last resort. This setup resulted in a list of 167 declension paradigms. In this data, declension class code (e.g., m[asculine]22, f[eminine]39) was adopted from Fronek's.

#### 2.2 Formal Concept Analysis: Method

To overcome the difficulties mentioned above, the current research adopts a somewhat unusual approach to the representation of paradigms. It gave up the idea of defining a paradigm in terms of word-forms. Instead, it explores the possibility of defining it in terms of relationships among word-forms. More specifically, the formal representation of a paradigm is equated with the set of identity relations among the declensions of a given word. Let me give you an example. Suppose we decided to describe the declension of *okno*, a neuter noun meaning "window." The declension it follows is the one given in Table 1. As exemplified in Tables 1 and 2, Czech nouns take 14 forms to reflect the combination of CASES (7 values) and NUMBERS (2 values).[3]

Next, we construct the **formal identity networks** over the combinations of all case forms. It is an encoding shown in Table 3, which is the Cartesian product of 14 forms × 14 forms minus 7 self-identity pairs and symmetrical pairs, yielding 91 values.

Building on the encoding scheme exemplified in Table 3, the declension of nouns were encoded to make

*Contact: `kow .k @ks .kyorin -u .ac .jp`

[1] Space limitation discouraged me to explain how declension is related to inflection, which is far from a trivial question.

[2] Note that *okno* and *místo* are both neuter nouns. Classification of declensional paradigms occur at multiple layers. The major divisions are made at the level of genders. Czech have three genders: MASCULINE, FEMININE and NEUTER and they realize qualitatively distinct paradigms at least in terms of word endings. Then minor divisions occur inside each given gender: there are different paradigms for neuter nouns, and the same is true of masculine and feminine nouns.

[3] There are a few exceptions to this. First, some nouns have only the singular declension. For example, *hovězí* (beef [neuter]) lacks the entire paradigm for plural forms. Second, some nouns have different stems for singular and plural paradigms.

Table 3: Matrix specifying 91 pairwise formal identities

| | sGen | sDat | sAcc | sVoc | sLoc | sIns | pNom | pGen | pDat | pAcc | pVoc | pLoc | pIns |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| sNom | sN=sG | sN=sD | sN=sA | sN=sV | sN=sL | sN=sI | sN=pN | sN=pG | sN=pD | sN=pA | sN=pV | sN=pL | sN=pI |
| sGen | – | sG=sD | sG=sA | sG=sV | sG=sL | sG=sI | sG=pN | sG=pG | sG=pD | sG=pA | sG=pV | sG=pL | sG=pI |
| sDat | – | – | sD=sA | sD=sV | sD=sL | sD=sI | sD=pN | sD=pG | sD=pD | sD=pA | sD=pV | sD=pL | sD=pI |
| sAcc | – | – | – | sA=sV | sA=sL | sA=sI | sA=pN | sA=pG | sA=pD | sA=pA | sA=pV | sA=pL | sA=pI |
| sVoc | – | – | – | – | sV=sL | sV=sI | sV=pN | sV=pG | sV=pD | sV=pA | sV=pV | sV=pL | sV=pI |
| sLoc | – | – | – | – | – | sL=sI | sL=pN | sL=pG | sL=pD | sL=pA | sL=pV | sL=pL | sL=pI |
| sIns | – | – | – | – | – | – | sI=pN | sI=pG | sI=pD | sI=pA | sI=pV | sI=pL | sI=pI |
| pNom | – | – | – | – | – | – | – | pN=pG | pN=pD | pN=pA | pN=pV | pN=pL | sN=pI |
| pGen | – | – | – | – | – | – | – | – | pG=pD | pG=pA | pG=pV | pG=pL | pG=pI |
| pDat | – | – | – | – | – | – | – | – | – | pD=pA | pD=pV | pD=pL | pD=pI |
| pAcc | – | – | – | – | – | – | – | – | – | – | pA=pV | pA=pL | pA=pI |
| pVoc | – | – | – | – | – | – | – | – | – | – | – | pV=pL | pV=pI |
| pLoc | – | – | – | – | – | – | – | – | – | – | – | – | pL=pI |

them formal context given to FCA.[4] For example, *okno* in Table 1 is encoded as the matrix in Table 4, and this matrix was then converted into a feature vector by serializing the distribution of values.

Table 4: Identify matrix of *okno*

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| – | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| – | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| – | – | – | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| – | – | – | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| – | – | – | – | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| – | – | – | – | – | – | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| – | – | – | – | – | – | – | 0 | 0 | 1 | 1 | 0 | 0 |
| – | – | – | – | – | – | – | – | 0 | 0 | 0 | 0 | 0 |
| – | – | – | – | – | – | – | – | – | 0 | 0 | 0 | 0 |
| – | – | – | – | – | – | – | – | – | – | 1 | 0 | 0 |
| – | – | – | – | – | – | – | – | – | – | – | 0 | 0 |
| – | – | – | – | – | – | – | – | – | – | – | – | 0 |

Some words, mostly frequent ones, show variations in declension. The noun *okno*, for example, has an alternative declension, where sDat and sLoc have the same form, *oknu*, as illustrated in Table 5. The identity matrix based method faces no trouble, however. This flexibility in dealing with variations is one of the greatest benefits of the encoding used in this study.

Table 5: Alternative declension of *okno*

| number | Nom. | Gen. | Dat. | Accu. | Voc. | Loc. | Inst. |
|---|---|---|---|---|---|---|---|
| SING. | *okno* | *okna* | ***oknu*** | *okno* | *okno* | ***oknu*** | *oknem* |
| PLUR. | *okna* | *oken* | *oknům* | *okna* | *okna* | *oknech* | *okny* |

### 2.2.1 Procedure for optimization

`Concept Explorer (ConExp) 1.3`[5] was used to execute FCA. This tool provides three functions useful for exploration: "attribute reduction," "object reduction," and "context reduction," through which redundant attributes/features and objects are discarded.These functions were employed at first to reduce the complexity of analysis, though all objects were retained later by undoing object reduction. Attribute reduction gave a set of 29 attributes.

To get optimal results, however, the heuristics in (1) was adopted:

(1) Avoid (i) attributes stating cross-categorial identity (e.g., sDat=pIns); and (ii) attributes mentioning Vocative (e.g,. sGen=sVoc).

After some investigation, however, it turned out that inclusion of sVoc=sNom and pVoc=pNom were useful and included. Thus, the set of relevant attributes was reduced to the one in (2) comprising 13 attributes, which is called the set of "legitimate" attributes:

(2) sNom=sAcc, sNom=sVoc, sGen=sDat, sGen=sAcc, sGen=sIns, sDat=sAcc, sDat=sLoc, sAcc=sIns, sLoc=sIns, pNom=pVoc, pGen=pAcc, pGen=pLoc, and pAcc=pIns.

Figure 1 specifies the Hasse diagram where the legitimate attributes in (2) are used. The encoding scheme used is the following: **m1∼m27** index masculine nouns; **f28∼f42** feminine nouns; **n49∼n57** neuter nouns; **p.{m, f, n, x}58∼p.{m, f, n, x}68** pronouns;[6] and **a69∼a70** adjectives. The integer after a gender class (e.g., m, f, n) is Fronek's class index; "**?**" is the class index if the data is taken from Kanazashi's database.

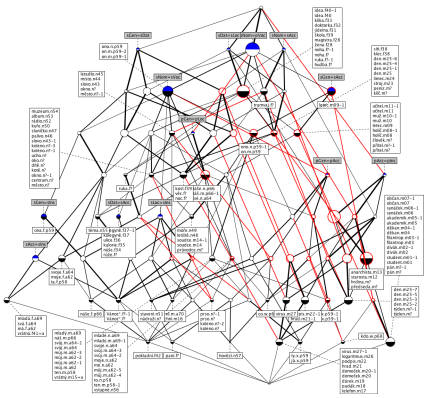

Figure 1: **FCA with all objects, with legitimate attributes**

The Hasse diagram in Figure 1 looks, admittedly, rather chaotic and is far from revealing, suggesting that optimization is necessary. Under this setting, the procedure in (3) was executed to achieve optimization:

(3) **Step 0.** Turn off all attributes. **Step 1.** Turn on a few (say (3∼5) attributes randomly. **Step 2.** Repeat the following until as many object separations and as few empty nodes as possible: Add one attribute at a time, and see if it yields no contradiction; if so, add one more attribute.

[4] Identity relations are not implemented in *F=G mainly because it requires more computation at least on the FCA tool used.

[5] Freely available at URL: `http://conexp.sourceforge.net/`

[6] Classes 60, 61, 63 and 65 are unused in this research.

In preliminary experimentations, inclusion of pronouns, on the one hand, and adjectives and adjectivals (i.e., nouns derived from adjectives), on the second, into FCA led to suboptimal results. Finally, the following facts were confirmed:

(4) a. The declensional classes for nouns and the one for pronouns/adjectives are different.

b. Feminine nouns are of a different kind within noun declensions.

Based on this, the results are to be presented separately.

There is another factor to complication that merits a mention here: in general, trade-off exists between the number of used attributes and the degrees of optimality of the produced FCA measured by the number of empty nodes in and the geometrical well-formedness of the resulting Hasse diagram. In short, the more attributes are used in FCA, the more empty nodes there are in the resulting Hasse diagram and the more complex geometry it has.

## 3 Results and Discussion

### 3.1 Result 1: FCA of pronouns and adjectives

To begin with, Figure 2 presents the FCA of pronoun/adjective declensions, where sLoc=sIns is discarded because it was proved to cause inconsistencies. The diagram suggest that relevant declensional classes are well separated and well classified. Remarkably, there are no gaps, i.e., empty nodes in the Hasse diagram. This means that this result is nearly optimal as far as we ignore the fact sLoc=sIns is not used.

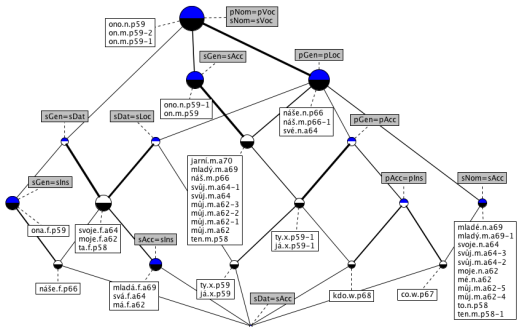


Figure 2: **FCA with pronouns and adjectives only**

Several things should merit mention. Note that attributes of the form F=G encode formal neutralization, and the best way to conceive how F=G works would be interpret it as violation of constraint *F=G to implement **anti-syncretism** in morphology.[7)]

In the Hasse diagrams, thus, the top (⊤) denotes an ideal situation where all possible cases of a paradigm have a unique realization, and the bottom (⊥) denotes an ideal situation where all cases have the same form. The situation corresponds to many languages that have no declension.

[7)] Admittedly, this is not a universal constraint.

### 3.2 Result 2: FCA of all data altogether

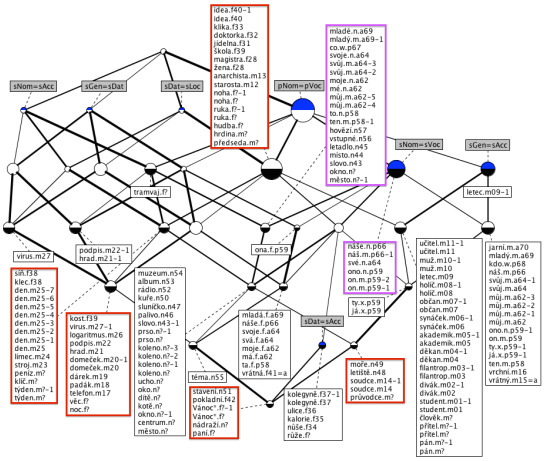


Figure 3: **FCA with all data with maximal number of attributes**

An FCA was attempted to see what we get if we try to implement as many attributes as possible to classify all declension classes while avoiding clashes. The result is shown in Figure 3 with the attributes in (5).

(5) sNom=sAcc, sNom=sVoc, sGen=sDat, sGen=sAcc, sDat=sAcc, sDat=sLoc, and pNom=pVoc,

Several things merit mention. Note first that the result in Figure 3 defines what is possible about Czech declensions under strict coherence. If there are nodes suffering from object overloading, they are unavoidable, indicating that **the system of Czech noun declension is not totally systematic by itself**, and involves unavoidable inconsistencies in it.

Note also that there are 6 empty nodes which are unused in the system. This means that the system is redundant, though necessary classification is achieved.

Note that variations, encoded by integer appendix like "-1," don't change declension class. This effect is not guaranteed to occur, and is not trivial.

Note that the following five sites (indicated by red edges, differentiated from the ones with magenta edges standing for anomalities involving pronominals) indicating inconsistencies in gender assignment:

(6) a. **f** {síň.f38, kleč.f38, tramvaj.f?} and **m** {den.m25-7, …, den.m25, týden.m?-1, týden.m?}

b. **f** {kost.f39, věc.f?, noc.f?} and **m** {virus.m27-1, logaritmus.m26, …, padák.m18, telefon.m17}

c. **f** { idea.f40-1, idea.f40, …, hudba.f? } and **m** { anarchista.m13, …, předseda.m?}

d. **n** {stavení.n51, nádraží.n?} and **f** { pokladní, …, paní.nf?}

e. **n** {moře.n49, letiště.n48} and **m** {soudce.m14-1, soudce.m14, průvodce.m?}

For (6a) and (6b), offenders are {síň.f38, kleč.f37} and {kost.f39, věc.f?, noc.f?}, which are semantically feminine at cost of violating a phonological generalization that words ending with consonants are feminines.

For (6c), offenders are {anarchista.m13, starosta.m12, hrdina.m?, předseda.m?}, which are semantically masculine at cost of violating a phonological generalization that words ending with *-a* are feminines.

For (6d), offenders are probably {pokladní.f42, Vánoc*.f1-1, Vánoc*.f?, paní.f?}, which are semantically feminine at cost of violating a phonological generalization that words ending with *-í* are neuter.

For (6e), it is somewhat hard to tell which is offending. Rather, it seems likely that this node is a true hybrid of neuters and masculines, since its direct licensers are a dominant class of neuter {muzeum.n54, album.n53, …, město.n?} and a dominant class of masculine {učitel.m11-1, …, muž.10-1, …, letec.m20, …}.

### 3.3 Result 3: FCA excluding offensive players

A note on terminology: if an optimization results in $n$-empty nodes, it is called "optimization at level $n$."

#### 3.3.1 Optimization at level 2

Other FCAs were attempted to get more aggressive optimizations. One of them is presented in Figure 4, constructed with the attributes in (7), by excluding feminine pronouns, adjectives and adjectives, but masculine and neuter adjectivals are included:

(7) sNom=sAcc, sNom=sVoc, sGen=sDat, sGen=sAcc, sGen=sIns, **sDat=sAcc**, sDat=sLoc, sAcc=sIns, and pNom=pVoc

sDat=sAcc is intentionally included, because this attribute, it turned out, serves as a defining feature to separate feminine declensions from other classes.

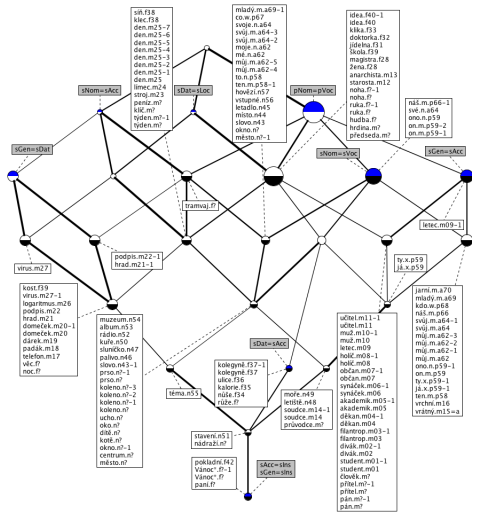

Figure 4: **FCA without feminine pronouns and adjectives**

Quite interestingly, the Hasse diagram in Figure 4 is geometrically well structured, nearly horizontally symmetrical. This suggests that **Czech declensional system is basically systematic as far as perturbation factors are carefully taken out**. Major candidate for perturbation is, as mentioned in (4), that pronouns and adjectives are a different kind, especially feminine ones.

#### 3.3.2 Optimization at level 0

Another FCA was generated to see if more data compression is possible, with pronouns, adjectives and adjectivals still excluded. The result is given in Figure 5, which is obtained with the following attributes.

(8) sNom=sAcc, sNom=sVoc, sGen=sDat, sGen=sAcc, sGen=sIns, sDat=sAcc, sAcc=sIns, and pNom=pVoc,

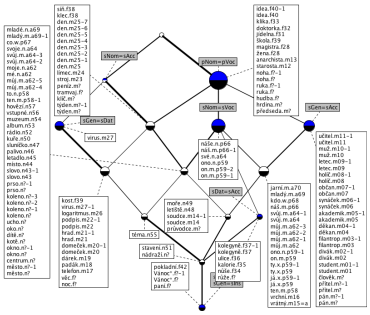

Figure 5: **FCA excluding feminine pronouns, adjectives, and adjectivals with more compression**

### 3.4 Discussion

One thing vital about the results is that the classification by FCA is not only descriptively satisfactory enough but also **unsupervised**, i.e., done without any linguistic knowledge (e.g., GENDER) that guides the output.

If the results above are adequate, they suggest that **declensional classes cannot be identified by simply specifying the concrete forms of a target word.** Second-order relation captured by identity matrix is necessary, though the psychological validity of this claim remains controversial. If valid, this means that the nature of knowledge by which inflectional paradigms come into existence is abstract in nature. Differently put, **a lemma is not a list of word forms**: rather, **a lemma is a structure, i.e., an organization over the inflectional relations among relevant forms**, of which the pairwise identity network is an essential part.

## 4 Concluding Remarks

This work examined the nature of inflectional classes by taking Czech declensional paradigms. Each class is represented as a network of pairwise identities. This representation scheme is shown to be effective in that FCA based on it captures relevant declensional classes successfully. The obtained results are promising but still insufficient in that direct comparison with similarity-based representation was not attempted yet, which future work will challenge.

## References


Fronek, J. (2010). *English-Czech Czech-English Compact Dictionary*. LEDA.

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

Kanazashi, K. (1998). *1500 Basic Words of Czech*. Daigaku Shorin. [Originally: 金指 久美子 (1998). チェコ語基礎 1500 語. 大学書林]].