**PROBLÉMY LEMMATIZACE A ZNAČKOVÁNÍ STARŠÍCH ČESKÝCH TEXTŮ**

V souvislosti s lemmatizací starších textů uvažujeme o následujících možnostech:

* **lemma** – v případech, kdy lexém je formálně totožný s dnešním, dostává běžné novočeské spisovné lemma, např. *vodou* : **voda** (kurzívou zde naznačujeme slovní tvar, tučně lemma); novočeské spisovné lemma dostává takový lexém i v případech, kdy se v paradigmatu vedle běžných dnešních tvarů vyskytují i starší/nářeční odlišné, jako např. *křídloma* (duál) : **křídlo***, prosich, prosi...* : **prosit**(aorist), *dělajiech,dělajieše...*: **dělat** (imperfektum)
* **hyperlemma** (formálně totožné s novočeským spisovným lemmatem) – užívá se

(a) v případech, kdy se starší nebo nářeční lexém formálně liší od odpovídajícího dnešního lexému provedením nebo neprovedením pravidelných hláskových změn v celém paradigmatu nebo jednotlivých tvarech; tyto lexémy dostávají spisovné novočeské lemma, např. *kóň/kuoň/kůň* : **kůň**; *bielý/bílý* : **bílý**; *dobrý/dobrej* : **dobrý**, *zelé/zelí*: **zelí**, *okolo/vokolo* : **okolo**...;

(b) v případech, kdy se lexém z období po roce 1850 liší způsobem psaní od odpovídajícího dnešního lexému (*jalousie* : **žaluzie**, *grandiosní* : **grandiózní**, *sorganisovat* : **zorganizovat**)

* **seznam variant** (zobrazovaný v případech, kdy lemma má více formálně blízkých podob) – předpoklad: uživatel by byl upozorněn na existenci formálně blízkých lexémů a mohl by si zvolit, zda chce vyhledávat pouze tvary lemmatu, které zadal, anebo tvary všech zobrazených formálně blízkých lemmat; prototypicky by mělo jít

(a) o případy, kdy je starší lexém formálně blízký dnešnímu, ale nejde o rozdíly uvedené zde výše u hyperlemmatu (uplatnění pravidelných hláskových změn a pravopisné rozdíly po r. 1850); např. *anděl/anjel, mlha/mha, pamlska/pamlsek*

(b) o případy, kdy v současné nebo ve starší češtině koexistují 2, 3... víceméně rovnocenné formálně rozdílné podoby lemmatu; např. *kolébka/kolíbka, moždíř/hmoždíř, -ismus/-izmus, dohmat/dochmat*

* **dobové lemma** (v nejnovější doložené podobě)– užívá se v případech, kdy starší lexém zanikl, např. stč. *šlachtata* : **šlachtata**

**K diskusi na úterním semináři 18. 11. si, prosím, utvořte názor na následující problémy:**

1. Jak lemmatizovat případy, kdy se v paradigmatu jinak shodném s novočeským vyskytuje odlišný tvar nominativu sg. nebo infinitivu, tj. tvar, který se standardně užívá jako lemma; tento tvar se v paradigmatu objevuje u sloves jen jednou, v infinitivu (*dát – dáti*), u substantiv obvykle dvakrát, v nominativu a akuzativu (*socialism – socialismus, tíž – tíže*); jde přitom o podoby lemmatu, podle nichž dnes nejspíš nebude nikdo vyhledávat (*dáti, socialism*, *tíž*)?
2. Jak lemmatizovat případy shodné s bodem 1 s tím rozdílem, že i dnes pravděpodobně bude někdo vyhledávat podle obou podob lemmatu: *klér – klérus, cykl – cyklus, zem – země*?
3. Jak řešit případy *cirk – cirkus*?
4. Lze řešit stejným způsobem (užitím hyperlemmatu, viz výše) všechny případy rozdílů vzniklých různým uplatněním pravidelných hláskových změn (např. *kolébka – kolíbka, řéci – říci, unést – unýst, mléko – mlíko, okénko – okýnko*)?
5. Jak lemmatizovat nepravidelné hláskové rozdíly v podobách slov, např. *sekera – sekyra, kralomoc – králomoc, knoflikář – knoflíkář, čtyrbřitý – čtyřbřitý, stkvít – skvít, vlaštovka – vlašťovka – lašťovka – laštovka, odbírat – odebírat, umístitel – uměstitel, sběratel – sbíratel, zaměřit – zamířit, knír – kňár, domněnka – důměnka, porcelán – porculán, závírka – zavírka – závěrka, nesený – nešený...*?
6. Jak lemmatizovat méně nápadné rozdíly svou podstatou slovotvorné, např.  *umístit – umístnit, chladomilný – chladnomilný, tedy – teda, nikdy – nikdý – nikda – nikdá, kamkoli – kamkoliv (– kamokoli – kamokoliv – kamkol), knihotisk – knihtisk, učeník – učenník – učedník – učedlník, lišta – lištna, virtuosný – virtuózní, měděrytina – měďorytina – mědirytina, dovědět se – dozvědět se, znít – vznít...*?
7. Jak lemmatizovat nápadnější slovotvorné rozdíly, např. *Ankara – Angora, Chorvatsko – Charvátsko – Hrvatsko – Chorvaty – Charváty, Regensburg – Řezno, Stuttgart – Stuttgard – Štutgart, Aristoteles – Aristotileš, Justina – Justýna, Čud – Čuchonec – Fin, Diokletian – Dioklecián, Assyrie – Asýrie, Edinburgh – Edinburk*

**Obecnější problémy:**

1. Jaký slovní druh uvádět v případech jako *po kazatelsku*, *ku předu, na levo...*?
2. Máme rozlišovat případy, kdy je trpné příčestí formálně shodné s jmenným tvarem adjektiva, např. *opuštěn – opuštěný, oběšen – oběšený, provrtán – provrtaný, broušen – broušený, probodnut – probodnutý, proklat – proklatý...*?
3. Máme rozlišovat mezi případy jako *dáti – dát, péci – péct* (v nichž existuje jistá, i když různě velká pravděpodobnost, že někdo zadá i méně běžnou/nespisovnou podobu lemmatu) a případy jako *vynést – vynest – vynésti – vynesti* (v nichž je zadání méně běžných podob lemmatu nepravděpodobné)? Máme-li je rozlišovat, tedy jak?
4. Máme užívat více lemmat v případech jako *načs, kdyžs, proň, bylť/byltě, mějmež*...?
5. Máme užívat několikanásobné lemma v případech nemožnosti rozhodnout, ke kterému lemmatu patří slovní tvar v případech jako *odřekli* : **odřeknout/odříci**;*vrhl* : **vrhnout/vrci**; *peku/peču* : **péci/péct;** genitiv sg. *země* : **země/zem**;dativ sg.*břichu*: **břicho/břich**; genitiv/dativ/lokál sg.*cirku*: **cirk/cirkus**... ? V případech jako např. *jako, tak…* (částice/adverbium/spojka…) by bylo třeba počítat s více lemmaty, ale i s více tagy.
6. Mají mít komparativy a superlativy samostatná lemmata (např. *menší* : **menší**,*přesněji*: **přesněji**), tj. nikoli lemmata prvního stupně (v daném případě **malý***,* ***přesný***)?