HiČKoK: History of Czech in Corpus Continuum

Daniel Zeman

December 9, 2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

HIČKOK

Historie češtiny
v korpusovém
kontinuu

Three-year project (September 2023 – November 2026)

T A
Č R

Funded by TAČR

Carried out by

ČESKÝ NÁRODNÍ KORPUS

ÚFAL

Národní knihovna
České republiky
National Library
of the Czech Republic

HIČKOK

Historie češtiny
v korpusovém
kontinuu

Three-year project (September 2023 – November 2026)

T A
Č R

Funded by TAČR

Carried out by

ČESKÝ NÁRODNÍ KORPUS

ÚFAL

Národní knihovna
České republiky
National Library
of the Czech Republic

Outline

Goals

zalakovaná , za což jsi velice vděčný . A hopla	,/,/Z:---------- dva/dva/ClIP1---------- ,/,/Z:----------	tři , čtyři . Vojáci už se postavili na nohy
jednoho jsou tak akorát pro hlodavce , na chodbě se	vyhnou/vyhnout/VB-P---3P-AA---P dva/dva/ClMP1---------- chodci/chodec/NNMP1-----A-----	. Pluješ po linoleu do Oddělení verifikace
to máš nechat plavat . Zatímco jsi byl pryč ,	brala/brát/VpFS----R-AA----I dva/dva/ClIP4---------- vzkazy/vzkaz/NNIP4-----A-----	, jeden od jakéhosi monsieura z oddělení čehosi a
měla barvy pšenice , aspoň tak sis to představoval ;	za/za/RR--4---------- dva/dva/ClIP4---------- měsíce/měsíc/NNIP4-----A-----	v Kansasu jsi ještě žádnou pšenici neviděl . Většin
, že by sis ji mohl s Dolly spojovat .	O/o/RR--4---------- dva/dva/ClIP4---------- roky/rok/NNIP4-----A-----	později byla Dolly pozvaná na svatbu na východě .
je svým lehkomyslným způsobem velkorysý . " Máte k sobě	vy/vy/PP-P1--2-------- dva/dva/ClIP1---------- hodně/hodně/Dg-------1A-----	blízko , " chceš vědět . " Podle mě je
tam našla chyby ? " Má seznam pěkně po ruce	:/:/Z:---------- dva/dva/ClIP4---------- opačné/opačný/AAIP4----1A-----	přízvuky , volební obvod ve střední Francii nesprávn
Alex Hardy . Vážně pokývne hlavou , jako byste právě	vy/vy/PP-P1--2-------- dva/dva/ClIP4---------- byli/být/VpMP----R-AA---I	osobami , které hodlá najít o půlnoci v Kláfině kanc
peníze z pojistky - jisté je jen , že zmizely	za/za/RR--4---------- dva/dva/ClIP4---------- týdny/týden/NNIP4-----A-----	. U vchodu vysoká žena se stříbrnými vlas
stříbrnými vlasy pečlivě zkoumá tvou pozvánku . Po stranách dveří	stojí/stát/VB-P---3P-AA---B dva/dva/ClMP1---------- ohromní/ohromný/AAMP1----1A-----	černoši v turbanech a paže mají zkřížené na prsou
paměti a pátráš po nějakých stopách . Ve vrchním šuplíku	objevíš/objevit/VB-S---2P-AA---P dva/dva/ClIP4---------- prázdné/prázdný/AAIP4----1A-----	obdélníkové balíčky . Jeden z nich vlastně úplně pr
. Kreditní kartou ho vyškrábneš na stůl a kartou také	uděláš/udělat/VB-S---2P-AA---P dva/dva/ClIP4---------- rovné/rovný/AAFP4----1A-----	řádky . Podíváš se na Megan . Čte . Klidně
, že má rozkošný zadeček . Dělal jsi s ní	skoro/skoro/Db---------- dva/dva/ClIP4---------- roky/rok/NNIP4-----A-----	a vůbec sis té nádhery nevšiml . Kolik jí vlastně
je velká a lesklá a je na ní mladší Megan	a/a/J^---------- dva/dva/ClMP1---------- muži/muž/NNMP1-----A-----	na jevišti . " To byla moje poslední role .
ten česnek , " zeptá se . Podařilo se ti	oloupat/oloupat/Vf--------A----P dva/dva/ClIP4---------- stroužky/stroužek/NNIP4-----A-----	. Jsou pěkně nahaté . " Moc výkonný zrov
radši natáhl . Zabořit se hlavou Megan do klína a	týden/týden/NNIS4-----A----- dva/dva/ClIP1---------- tak/tak/Db----------	zůstat . Postel je jen pár kroků odsud . Nakloníš
že věkem bude ještě pod zákonem . Pod očima si	namalovala/namalovat/VpFS----R-AA---P dva/dva/ClIP4---------- purpurové/purpurový/AAIP4----1A-----	pr…, …, …připomínat lícní kosti . Už víš
mu zatřeseš ramenem . Zaboří se obličejem do polštáře .	Uděláš/udělat/VB-S---2P-AA---P dva/dva/ClIP4---------- řádky/řádek/NNIP4-----A-----	a sedneš si do křesla . Dnes je to přesně
" Je to pravda . Musel jsem s těma numerama	bejt/být/Vf--------A---6I dva/dva/ClIP1---------- večery/večer/NNIP1-----A-----	za sebou a samozřejmě , že expres Allagash vykol
. " V koutě podlaží je bokem několik pokojíků .	První/první/CrMP1---------- dva/dva/ClIP1---------- jsou/být/VB-P---3P-AA---I	plné kokainových fandů a seriózních žvanilů . Třetí
zotročit , že myslí i na jiné věci . Ukrojila	ti/ty/PH-S3--2-------- dva/dva/ClIP4---------- tlusté/tlustý/AAIP4----1A-----	krajíce . Kůrku měly spálenou , ale uvnitř byly navlh
li nyní začít svou zpověď , musím si p		, které mám obzvlášť rád . " Velikou , tajemnou
dobré . Hříchů má špatná vláda mnoho , a		: pošlapání Spravedlnosti a vláda Tyranie . Ty ostat
jsem tam mnohokrát a pokaždé jsem si odtamtud		dojmy : hlučný sobotní trh , rozpínající se po městě
a to právě v papežském paláci . Stanuly		, filoanjouovský , který navrhoval francouzského ka
kulatý stůl v kolbiště prudkých střetů . N		, a stále ještě nebylo vidět světlo habemus papam

zůstat/Vf--------A----P

▲ Knoflík na dveřích se otáčí sem a tam . Allagash tě šťouchá do ruky a rty se horečně vyptává . Zámek cvakne a dveře se rozletí . Stojí v nich Alex Hardy . Vážně pokývne hlavou , jako byste právě vy dva byli osobami , které hodlá najít o půlnoci v Kláfině kanceláři . Snažíš se rychle vymyslet povídačku , kterou by spolkl . Tad třímá dřevěný metr , který našel za dveřmi . " Ty jsi nás , Alexi , vylekal . ▼

Výskytů: **498 777** | i.p.m.: **965,12** (vztaženo k celému korpusu) | ARF: **277 963,99** | Výsledek je setříděn

Výběr řádků: základní

1 / 12 470

Inga Ábele Ostřice Hra o	dvou/NUM/Case=Loc|Number=Plur|NumForm=Word|NumType=Card	dějstvích Osoby Vilhelms Putns Lija , jeho že
zmlkne .) A tak dál a tak dál .	Dvě/NUM/Case=Nom|Gender=Fem|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	stránky a fotografie . Všichni to čtou , celé m
vinář ke mně přišel s lahví koňaku , že jsem	dva/NUM/Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	dny nejedla … , hodně toho tam není napsáno
. Lija . Aleksisi , přečti ještě něco . Ti	dva/NUM/Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	už mě omrzeli . leva přistoupí ke stolu a sebo
Vilhelms . Teď je to snad v pořádku . Už	dva/NUM/Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	roky je pryč . Teď si ho sem přivolala maminl
děje jenom v naší čtvrti . Lija . Měla jsem	dva/NUM/Animacy=Anim|Case=Acc|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	syny – vlastního Aleksise a ještě jednoho . V
Aleksis . To je teda vážně ostuda . A já	dva/NUM/Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	roky žil jak anděl . Santa (Usměje se ,
knem do tmy . Hledá cigarety , ale najde jen	dvě /NUM/Case=Acc|Gender=Fem|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	prázdné krabičky . Zapíná rádio . Přelaďuje s
… Ale neřekla jsem ti , že ve skutečnosti už	dva/NUM/Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	měsíce žiju s jiným týpkem , jo , no ,
i vyšťourával zbytky sena . Po jednom či po	dvou/NUM/Case=Loc|Number=Plur|NumForm=Word|NumType=Card	se děti vracely zpět k autu a sypaly otci do
oční stisk čelistí . Na břehu ležely na dekách	dvě/NUM/Case=Nom|Gender=Fem|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	tělnaté ženy a vítaly oba příchozí . U nikoho
dcera a syn , žena , muž a sestra ,	dvě/NUM/Case=Nom|Gender=Fem|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	sestry s bratrem nebo otec s dvěma dcerami
a sestra , dvě sestry s bratrem nebo otec s	dvěma/NUM/Animacy=Inan|Case=Ins|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	dcerami . Jejich radost a výkřiky byly urputné
lo malého města , vráščitá , unavená žena s	dvěma/NUM/Animacy=Inan|Case=Ins|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	nevychovanými syny . Pracovala jako pošťač
mlčky . Bertolds udělal jeden krok , chlapec	dva/NUM/Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	. V každé ruce se mu pohupoval sandál . Za
ejme , že jsem ještě nic nepropásl . Koupím	dvě/NUM/Case=Acc|Gender=Neut|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	pořádná kola , stan a vezmu si dovolenou . F
. Spěchaly dál přes hřbitov . U vchodu , kde	dvě/NUM/Case=Nom|Gender=Fem|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	borovice připomínaly tajemný vstup na onen
Prohodila to jen tak , mezi řečí , možná mezi	dvěma/NUM/Animacy=Inan|Case=Ins|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	sklenkami vína . Ale moje fantazie už začala
jet do Buenos Aires . , K tomu musej bejt	dva/NUM/Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	, ‘ přerušil ji vždycky muž , který tu teď
ýho všimnul a ocenil ho , k tomu musej bejt	dva/NUM/Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur|NumForm=Word|NumType=Card|NumValue=1|NumValue=2|NumValue=3	. ~ # “~ Regína se odmlčí , nanovo se
če pozoruje každý její pohyb . Bílá skvrna se	dvěma/NUM/An	žhnoucíma , dehtově černýma očima na vzor
Nechce se jim bez ní odejít . Pokládají ji na	dvě/NU	suché smrkové větve tak , aby se očními otvo
ěd očima . Nazelenalá kost orámuje krajinu	dvěma/NUM/An	okny . Jsou tak široká , že se musí dívat
sklem auta , tento spánek vyvolal vizi o nich		, o ní a Pávilsovi v prvotním lese , ukázal
chalupu . Potraviny spravedlivě rozdělili do	dvou/NUM/Anim	baťohů , jen dočerna vyuzený kanič hřbet vza
ela si s obavami leva a nespyšela žádné raz	dva/NUM/Anim	tři . Ve světle reflektorů viděla Pávilsův pro

Výchozí zobrazení | Formátovaný text

▲ dobře možné , že se ze slona narodil První Sníh ! Dino ! Vlk vyjde ven . Dveře jsou otevřené . Silně prší . Po schodech schází Aleksis a vyhlíží oknem do tmy . Hledá cigarety , ale najde jen dvě prázdné krabičky . Zapíná rádio . Přelaďuje stanice . Vypíná rádio . Opravuje kresbu . Sedne si ke stolu . Lehne si na pohovku . Přistoupí k oknu , protože přijíždí auto . Reflektory ho na moment osvětlí a auto ▼

1492	mi se , že jest . V tom městě byli	**sme dva dní**	, chleba vody i rozličného vovoce bylo co dosti kúpiti
1492	i na vodu , neb do něho teče sedmdesát potokóv	**a dva .**	Pak na kupecké věci jest převelmi bohaté , ješto na
1492	. Pak opět dne jiného ráno vstavše , šli se	**mnú dva bosáci**	do Jeruzaléma , neb hora Sion jest od Jeruzaléma jako
1492	kus zdi kostelnie , a tak mi pravili , že	**zabilo dva černé**	křesťany , a nedadie jim toho zase opraviti . V
1492	ješto držie křest , světie toliko neděli , ti mají	**jedno dva strychy**	nebo znamenie na tváři . Třetí toliko držie křest ,
1492	a některé muož tlušče býti . Vinic nekopají , ale	**zapřaha dva buvoly**	do pluhu i obvorávají je na všecky strany , totižto
1492	vašeho . V Gaze byli sme týden . Odtud sme	**jeli dva dní**	i přijeli sme k jednomu městečku a neviděli sme ho
1492	daktylóv jako chmele , že jich z toho kolečka natrhá	**se dva vĕrtele**	a z druhého viec . A na jednom dřevě bude
1492	toho voda z trub teče . A u toho koryta	**sedíta dva starci**	a ktož koli přijde k té mřéži a chce píti
1492	, a když sem toho pilen byl , potom mi	**poručil dva a**	tak dále , až mi i jiné věci porúčel .
1492	byla , tehdy na osle jede a okolo nie jdú	**služebníci dva nebo**	třie a tvář jejé jest všecka zakryta čistým tafartem černým
1492	a netoliko já , ale všichní kupcí . Byla se	**nám dva pacholky**	roznemohla , kteréžto tajně mezi sebú chovali sme a nemocné
1835	ještě vypravoval , kdyby ho příchozí nebyli vytrhli . Byli	**to dva cikáni**	. Napřed šel malý , po maďarsku vystrojený mladík ,
1835	Mladý cikán postoupil , spatřiv vysloužence a ostatní , zaražený	**as dva kroky**	zpět , jako by byl někoho jiného očekával ; opět
1835	mezi křovím nějakou rozsedlinou dolů . Starší lezl za ním	**asi dva sáhy**	hluboko , a oba stanuli v skalnaté kotlině asi dva
1835	dva sáhy hluboko , a oba stanuli v skalnaté kotlině	**asi dva sáhy**	dlouhé a sáh široké . Kolem nich strměla písečná skála
1835	se vám vracím jednou domů ; jest to ode dneška	**za dva měsíce**	osm let , já jsem si to zapsal . -
1835	a ještě jeden šli před nimi s rozžatými pochodněmi .	**Ostatní dva ,**	pomohše jim mrtvol přendati přes kostničí kámen , šli ozbrojeni
1835	prsa a zacláněl hlubokou ránu jejich . Zachvěl se cikán	**; dva rychlé**	kroky přistoupil blíže a po jeho nad hlavu vyzdvižené pravici
1835	tlačilo nesčíslné množství úkolního lidu . Na nejvyšším jeho vrcholku	**stály dva mladé**	stromky , uprostřed nich strměl čekan a kolem toho se
1835	strana kněz , po druhé mistr poprauí . Za ním	**vedli dva pacholci**	bláznivou Angelínu uvázanou na dlouhém provaze , která do výšky
1792			přišli , kterouž ona podlé
1792			zlatý Radoji , jak
1869			výtiscích ještě na skladě . Belletristický
1869			
1869			

▲ na palmě , neb když z daleka na ně hledí , zdá se jako voháňka , pak v tom větvěví sú kola vyrostlá jako melé a na tom melí bude daktylóv jako chmele , že jich z toho kolečka natrhá se dva vĕrtele a z druhého viec . A na jednom dřevě bude takových kol čtyři nebo pět . Také v tom městečku zdá mi se , že největšie živnost jich jest daktyle , neb odtud dodávají daktylóv do všech zemí okolních kupcí ▼

Vokabulář webový

Webové hnízdo pramenů k poznání historické češtiny

Co je VW | Aktuality | Slovníky staré češtiny | Korpusy | Edice | Mluvnice | Digitalizované slovníky | Odborná literatura | Audioknihy | Zdroje | Nástroje | Kontakty a odkazy | Připomínky

dva Hledej

ptáci k lepu ; sto jich pobitých odvlekú a cis'úc jich pod zed tekú . Takž ten šturm stá ješče viece nežli pilná	dva	měsiecě ; dnem i nocú bez přěstánie nódpočinu šturmovánie ; jakž obirtli tako praky , nódpočinúce čas vša
I řěch : „ Už … pamět jich . … učinil toto … . “ … múdrosti … skonánie prozřěli . … honieše jeden tisúc a	dva	… desět tisúcóv . Však proto , že bóh … prodal jě a hospodin zaklopil jě . Nenie vedě bóh náš jako bohov
ověčieho těla jsa . Rovný otcu podlé božstvie , menší otcě podlé člověčstvie . Jenže ač bóh jest a člověk , ne	dva	však , ale jeden jest Kristus . Jeden zajisté ne obrácením božstvie v tělo , ale přijětím člověčstvie u bozě .
řiven , buď s tobú vet ! “ Pakliť sě lépe rozmyslí , ješčeť lepší čin vymyslí , jímžtoť chlapa neochudí , avšak	dva	voly vylúdí . Jako byl jeden učinil vládyka , že byl vylúdil dva voly , s hnědým plavého , u svého kmetě d
in vymyslí , jímžtoť chlapa neochudí , avšak dva voly vylúdí . Jako byl jeden učinil vládyka , že byl vylúdil	dva	voly , s hnědým plavého , u svého kmetě dobrého , jehož obviniti chtieše , ale sličně nemožieše . Pozvav j
omínaj , také mně něco z diela daj ! “ „ Nejmám , pane , nice toho , rozdal sem málo i mnoho ; kromě ješče	dva	voly jmám , oba na ořěřě prodám , na vosku a na příkrově , v němžto mě pohřebú v rově . “ „ Ješče máš d
vás jiní . A když jima to vše povědě , již na smrtnej posteli sedě , kakž kolivěk mdle dycháše , proti diáblu	dva	dni bojováše , prosě za vše své milé syny , aby Bóh spustil jich viny , i za jiné , ktož budú jeho slúhy , aby
j… … přiblíži … ; … jeho ne… ; … k němu krátcí … , … jeho všichni … . A když sě jako ze sna probudi ,	dva	a sedmdesát sě … …ocí vzbudi ; smrt všed… na jeho hrudi , hned z něho duši vypudi . Tak sě on svých hi
edno pláčě jej zbývášě , protož ústavně plakášě ; a když sě plačíc nakloni a prozřěvši v hrob sě vkloni , uzře	dva	anděly z nebe u bielém rúšě blíz sebe , vniž k hlavě a k nohám viece , nevzdál ot sebe sediece . Otázasta ji
ám , k dolóm i k skalám , prosiece jich , aby je skryly , na ně padnúce , přěd mě jíti nesmějíce . A v tu dobu	dva	zloděje s ním vediechu , nad nimažto popraviti chtiechu . Když Ježúšě vyvedú z města na popravu , tehdy
Ježíš , vecě k ňemu : „ Šimone , jmám tobě cos pověděti . “ A on vecě : „ Mistře , pověz ! “ I vecě Ježíš : „	Dva	biešta dlužna jednomu lichevníku , jeden bieše dlužen patset peněz a druhý patdesát . A když nejmějiešta c
modliti . I stala sě jest krása jeho obličějě jinaká , když sě modlěše , a jeho rúcho bielé stkvieše sě . A tehdy	dva	mužě mluviešta s ním . A to biešta Mojžieš a Heliáš , viděna u veleslavenství , a mluviešta o jeho příhodě
lémě . Ale Petr a ta , ješto s ním jdiešta , otrápeni biechu snem . A procítivše uzřěchu jeho veleslavenstvie a	dva	mužě , ješto s ním stojiešta . I stalo sě jest , když otjidešta ot ňeho ta dva mužě , vecě Petr k Ježíšěvi : „ Př
ivše uzřěchu jeho veleslavenstvie a dva mužě , ješto s ním stojiešta . I stalo sě jest , když otjidešta ot ňeho ta	dva	mužě , vecě Petr k Ježíšěvi : „ Přikazateli , dobré jest nám tuto býti . A učiňme tři stany : jeden tobě a jede
ú ruku ku pluhu a ozřě sě za sě , nebude klčen k království božiemu . “ Pak potom znamena hospodin jiných	dva	a sedmdesát i rozesla jě po dvú do každého města všudy , kamž jmějieše sám přijíti . I mluvieše jim : „ Jis
yší , mě slyší , a kto vámi hrdá , mnú hrdá , a kto mnú hrdá , hrdá tiem , kto mě jest poslal . “ I vrátilo sě jest	dva	a sedmdesát učedlníkuov s radostí řkúce : „ Hospodine , také i běsové jsú nám poddáni ve jmě tvé . “ I vec

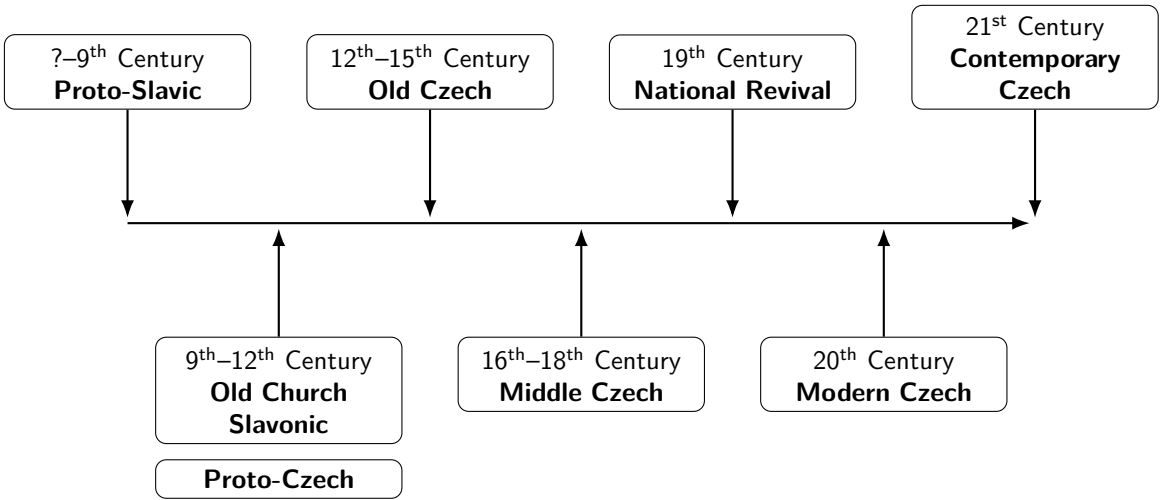
doc Bible drážďanská, Lukášovo evangelium; 2. polovina 60. let 14. století; rukopis; Německo; Drážďany; Saská státní knihovna – Státní a univerzitní knihovna v Drážďanech (Die Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden); Mscr.Dresd.Oe.85; 14; Bibl; BiblDrážď; próza; biblický text; ediční poznámka; Hlaváčová Svobodová, Andrea (Ústav pro jazyk český AV ČR, v. v. i.) – Voleková, Kateřina (Ústav pro jazyk český AV ČR, v. v. i.)

Goal

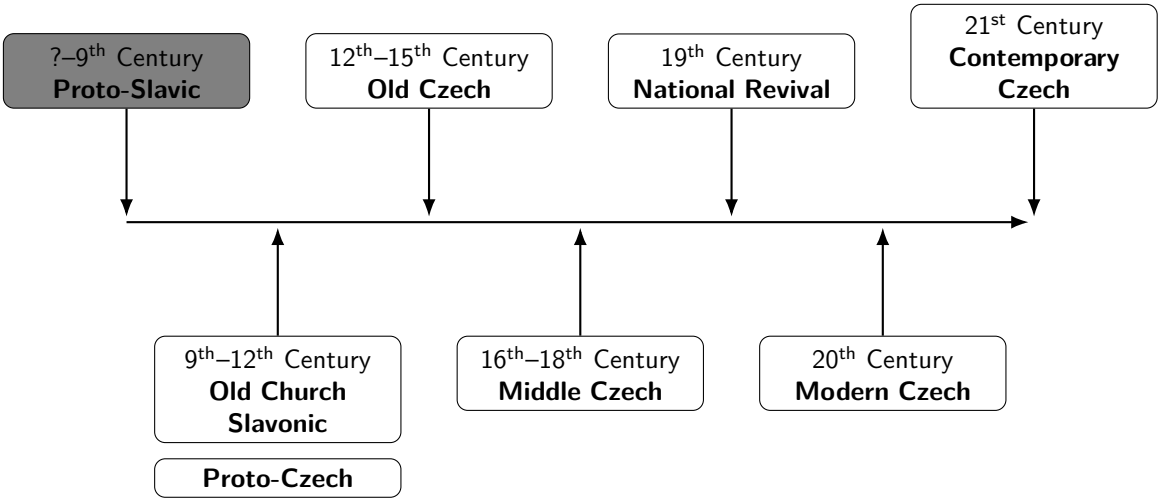
- Manual annotation of texts from all stages of Czech
- Universal Dependencies annotation style
- Only morphology (lemma, UPOS, features); no syntax
- Large enough to train a model ('etalon', ca. 100K words per period)
- UDPipe models will be freely available
  - Unfortunately, we are not allowed to make the data available for download

- Sources of data: ÚJČ, ČNK, NK
  - For some stages, no large corpus was available so far
  - Not even unannotated

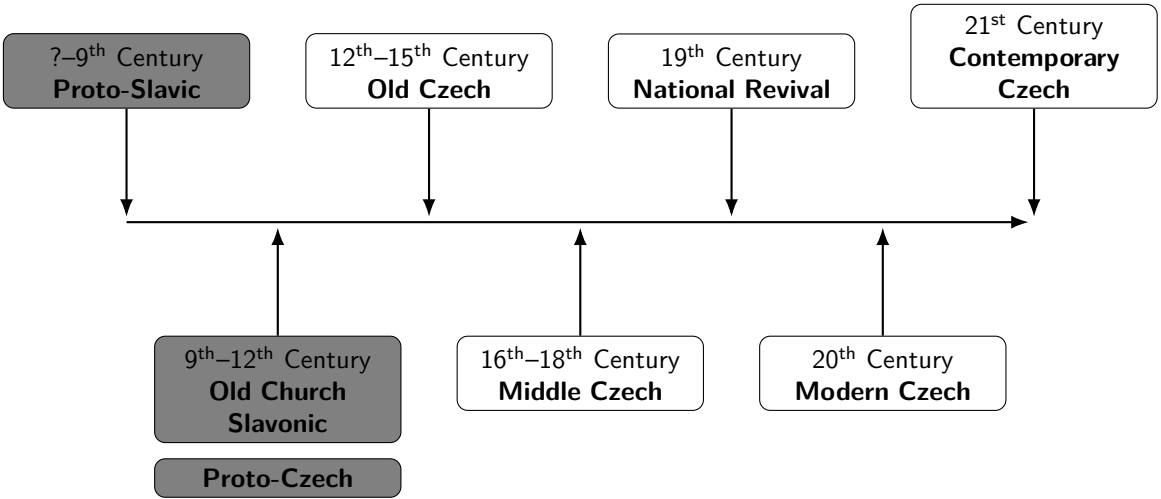
History of the Czech Language



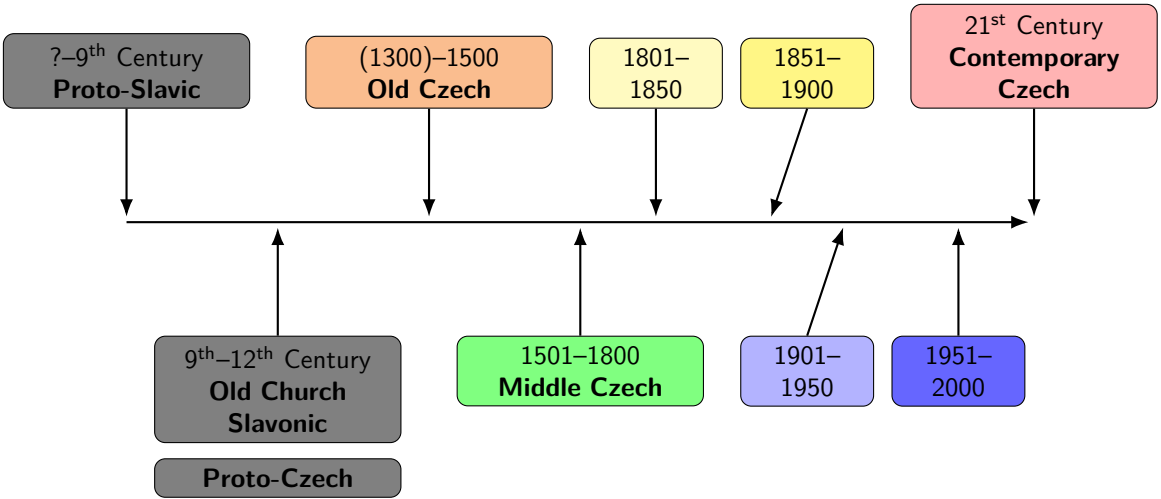
History of the Czech Language



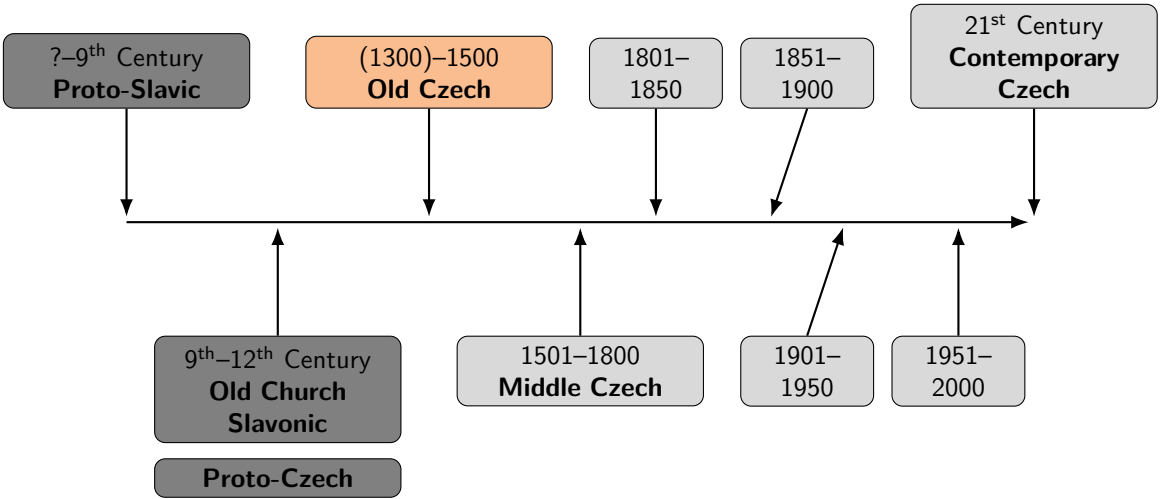
History of the Czech Language



History of the Czech Language



History of the Czech Language



Old Czech: Workflow

Old Czech Etalon

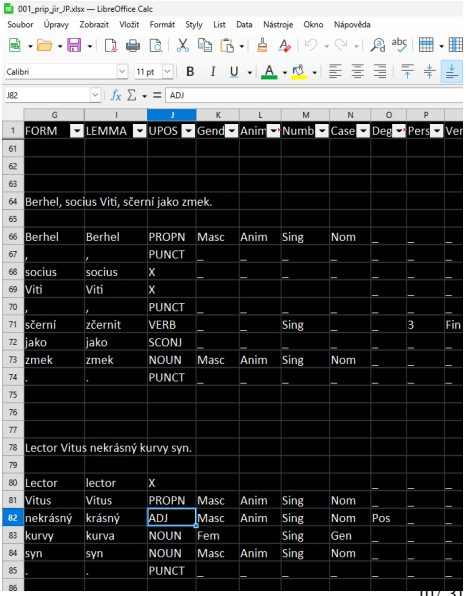
- 45 texts, oldest dated between 1290 and 1310
- Counts (may change during annotation):
  - 1294 sentences
  - 122,819 tokens
- Genres
  - Bible translations
  - psalms, prayers, sermons
  - epic
  - chronicles
  - regulations, law
  - medicine
  - proverbs, dream books
  - notes added to other texts

Workflow

- Preprocess by UDPipe trained on Modern Czech (PDT)
  - Plus rule-based script that fixes some frequent words (e.g. *řka* "saying")
  - $\Rightarrow$ sentence segmentation, lemmatization, UPOS, features, dependency relations

Workflow

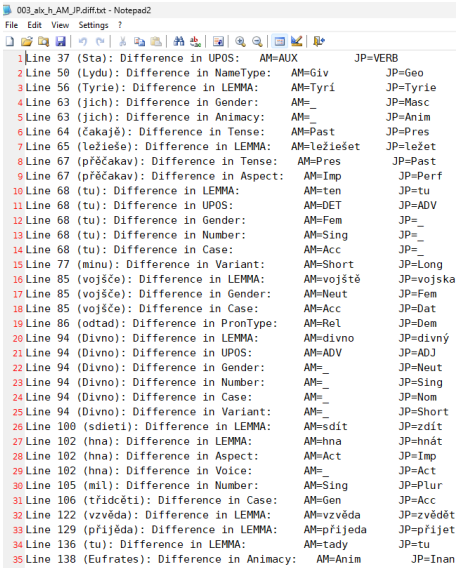
- Preprocess by UDPipe trained on Modern Czech (PDT)
  - Plus rule-based script that fixes some frequent words (e.g. *řka* "saying")
  - ⇒ sentence segmentation, lemmatization, UPOS, features, dependency relations
- Convert to TSV (table)
- Annotate in Excel



	G	I	J	K	L	M	N	O	P	
1	FORM	LEMMA	UPOS	Gend	Anim	Numb	Case	Deg	Pers	Ver
61										
62										
63										
64	Berhel, socius Viti, sčerní jako zmek.									
65										
66	Berhel	Berhel	PROPN	Masc	Anim	Sing	Nom	_	_	_
67	,	,	PUNCT	_	_	_	_	_	_	_
68	socius	socius	X	_	_	_	_	_	_	_
69	Viti	Viti	X	_	_	_	_	_	_	_
70	,	,	PUNCT	_	_	_	_	_	_	_
71	sčerní	zčernit	VERB	_	_	Sing	_	_	3	Fin
72	jako	jako	SCONJ	_	_	_	_	_	_	_
73	zmek	zmek	NOUN	Masc	Anim	Sing	Nom	_	_	_
74	.	.	PUNCT	_	_	_	_	_	_	_
75										
76										
77										
78	Lector Vitus nekrásný kurvy syn.									
79										
80	Lector	lector	X					_	_	_
81	Vitus	Vitus	PROPN	Masc	Anim	Sing	Nom	_	_	_
82	nekrásný	krásný	ADJ	Masc	Anim	Sing	Nom	Pos	_	_
83	kurvy	kurva	NOUN	Fem		Sing	Gen	_	_	_
84	syn	syn	NOUN	Masc	Anim	Sing	Nom	_	_	_
85	.	.	PUNCT	_	_	_	_	_	_	_
86										

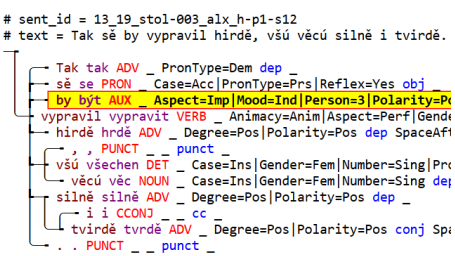
Workflow

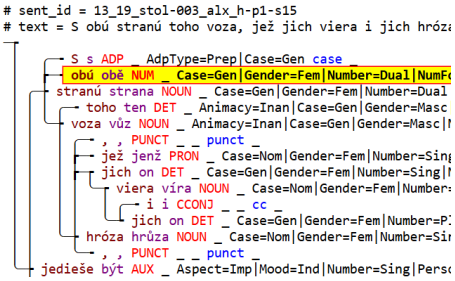
- Preprocess by UDPipe trained on Modern Czech (PDT)
  - Plus rule-based script that fixes some frequent words (e.g. *řka* "saying")
  - ⇒ sentence segmentation, lemmatization, UPOS, features, dependency relations
- Convert to TSV (table)
- Annotate in Excel
- Two annotators per file
- Convert to CoNLL-U, postprocess, diff
  - ⇒ one annotator resolves differences



Workflow

- Preprocess by UDPipe trained on Modern Czech (PDT)
  - Plus rule-based script that fixes some frequent words (e.g. *řka* "saying")
  - ⇒ sentence segmentation, lemmatization, UPOS, features, dependency relations
- Convert to TSV (table)
- Annotate in Excel
- Two annotators per file
- Convert to CoNLL-U, postprocess, diff
  - ⇒ one annotator resolves differences
- Official UD validation
- Additional consistency checks in Udapi
  - required & allowed features
- Return, fix, repeat...





Czech UD Treebanks

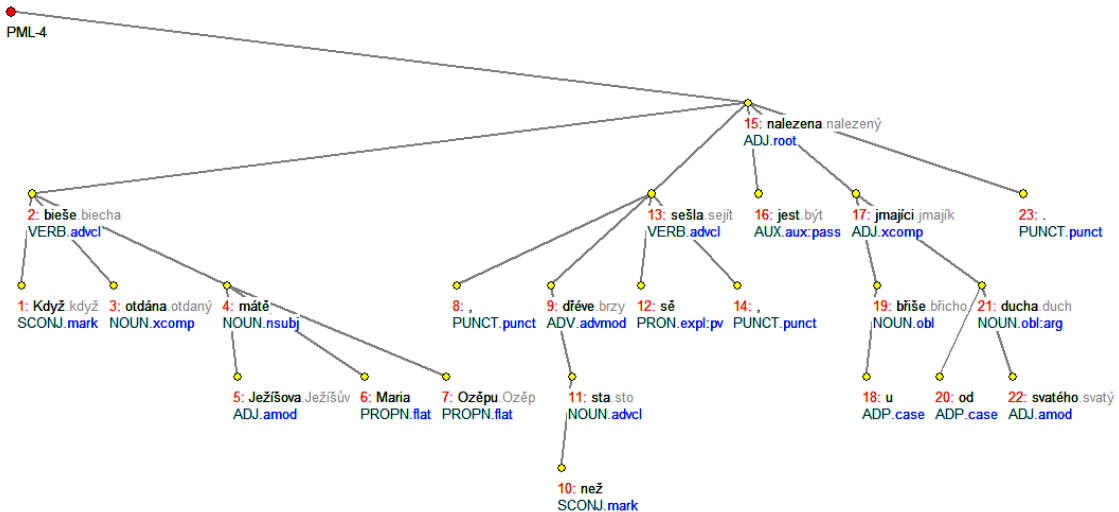
- PDT (Prague Dependency Treebank)
  - Lidové noviny + Mladá Fronta + ČM Profit + Vesmír, 1993–1994
  - 87K sentences, 1.5M words
- CAC (Czech Academic Corpus / Korpus věcného stylu)
  - non-fiction, 1971–1985
  - 24K sentences, 493K words
- FicTree
  - fiction, from Czech National Corpus, 1991–2007
  - 12K sentences, 166K words
- CLTT (Czech Legal Text Treebank)
  - The Accounting Act (Zákon o účetnictví), 1991–2016
  - 1K sentences, 36K words
- PUD (Parallel Universal Dependencies)
  - online news + Wikipedia, translated from en/de/fr/it/es, around 2016
  - 1K sentences, 18K words
- Poetry
  - poetry, 19^{th} century
  - 297 sentences, 6K words

PDT Model vs. Old Czech Data

- Genre, vocabulary: news vs. Bible
- Old vocabulary
- Orthography
  - Cleaned, transcribed, unified
  - But still not modern forms: *sě, viece*

- Grammar:
  - Dual number
  - Simple past (imperfect, aorist) *(bieše, vecě, jide)*
  - Converbs (přechodníky) *(řka, přistúpiv)*

Example Parse (UDPipe 2.0 on UD PDT 2.6)



Example Parse (UDPipe 2.0 on UD PDT 2.6)

ID	FORM	LEMMA	UPOS	XPOS	FEATS
1	Ale	ale	CCONJ	_	_
2	Kristovo	Kristův	ADJ	_	Case=Nom|Gender=Neut|Gender[psor]=Masc|NameType=Sur|Number=Sing|Poss=
3	porozenie	porozenie	NOUN	_	Case=Nom|Gender=Neut|Number=Sing|Polarity=Pos
4	tak	tak	ADV	_	PronType=Dem
5	bieše	biešat	VERB	_	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|Verl
6	.	.	PUNCT	_	_

Bootstrapping

- 10 texts already annotated
  - 958 sentences, 20K words

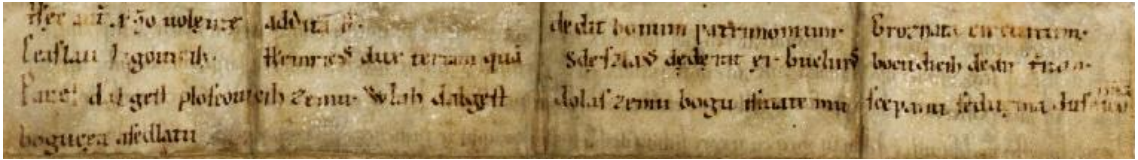
- Retrain UDPipe
- Reparse the remaining files
- Hopefully better ⇒ less work for annotators

- Earlier experiments with Dresden Bible, Gospel of Matthew:

Model	Lemma	UPOS	Features
UDPipe 1.2 on PDT UD 2.5	70	77	55
UDPipe 1.2 on FicTree 2.10 + 1669 words of Matthew	**78**	**85**	**65**

Old Czech: Interesting Issues

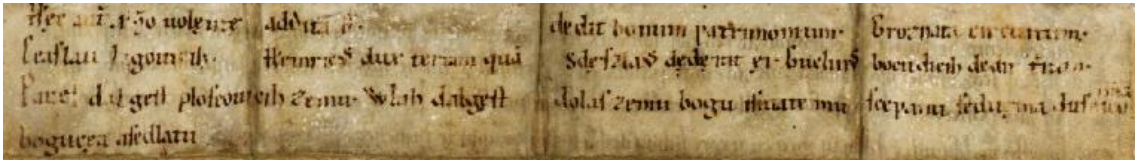
Modern Orthography vs. Modern Language



- *Pauel dal geſt ploſcouicih zemu Wlah dalgeſt dolaſ zemu bogu ıſuıatemu ſcepanu ſeduema duſnıcoma bogucea aſedlatu*
- *Pavel dal jest Ploskovicích zem'u, Vlach dal jest Dolas zem'u bogu i sv'atému Ščepánu se dvěma dušníkoma, Bogučeja a Sedlatu.*
- *Pavel dal v Ploskovicích zemi, Vlach dal v Dolanech zemi bohu i svatému Štěpánovi se dvěma dušníky, Bogučejem a Sedlatou.*

- "Pavel gave land in Ploskovice, Vlach gave land in Dolas to God and St. Stephen with two villeins, Bogučej and Sedlata."

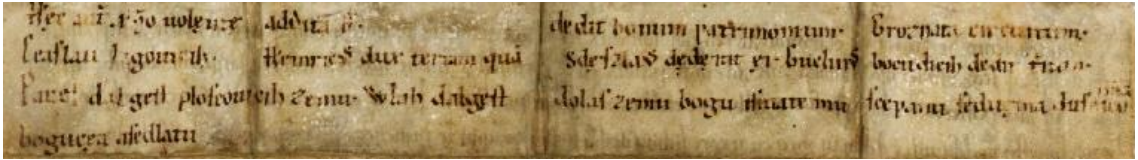
Modern Orthography vs. Modern Language



- *Pauel dal geſt ploſcouicih zemu Wlah dalgeſt dolaſ zemu bogu ıſuıatemu ſcepanu ſeduema duſnıcoma bogucea aſedlatu*
- *Pavel dal jest Ploskovicích zem'u, Vlach dal jest Dolas zem'u bogu i sv'atému Ščepánu se dvěma dušníkoma, Bogučeja a Sedlatu.*
- *Pavel dal v Ploskovicích zemi, Vlach dal v Dolanech zemi bohu i svatému Štěpánovi se dvěma dušníky, Bogučejem a Sedlatou.*

- "Pavel gave land in Ploskovice, Vlach gave land in Dolas to God and St. Stephen with two villeins, Bogučej and Sedlata."

Modern Orthography vs. Modern Language



- *Pauel dal geſt ploſcouicih zemu Wlah dalgeſt dolaſ zemu bogu ıſuıatemu ſcepanu ſeduema duſnıcoma bogucea aſedlatu*
- *Pavel dal jest Ploskovicích zem'u, Vlach dal jest Dolas zem'u bogu i sv'atému Ščepánu se dvěma dušníkoma, Bogučeja a Sedlatu.*
- *Pavel dal v Ploskovicích zemi, Vlach dal v Dolanech zemi bohu i svatému Štěpánovi se dvěma dušníky, Bogučejem a Sedlatou.*

- "Pavel gave land in Ploskovice, Vlach gave land in Dolas to God and St. Stephen with two villeins, Bogučej and Sedlata."

Tokenization

- UD concept of multiword tokens
  - one orthographic word → multiple morphosyntactic words

- Modern Czech examples:
  - *kdybychom* "if we would" = *když*/***SCONJ*** + *bychom*/***AUX***
  - *udělalas* "you.**Fem.Sing** did" = *udělala*/***VERB*** + *jsi*/***AUX***
  - *ses* = *jsi*/***AUX*** + *se*/***PRON***

Multiword Tokens in Old Czech

Preposition + accusative pronoun:

nač	=	*na*	*co*	**PronType=Rel**
naň	=	*na*	*něj*	**PronType=Prs**
naňž	=	*na*	*nějž*	**PronType=Rel**
oč	=	*o*	*co*	**PronType=Rel**
oň	=	*o*	*něj*	**PronType=Prs**
oňž	=	*o*	*nějž*	**PronType=Rel**
...				
přědeň	=	*před*	*něj*	**PronType=Prs**
skirzěňž	=	*skirzě*	*nějž*	**PronType=Rel**
...				
zaňž	=	*za*	*nějž*	**PronType=Rel**

The Particle-Clitic *-ť*

Function hard to nail down: discourse connective, emphasis, sometimes even personal pronoun *ti* "you.**Dat**".

mnohémuť = *mnohému ť*

Examples:

- *pakť, tedyť, ještěť, ktožť, mnohémuť, vzpomenúť, neopustíť...*

Except for lexicalized cases:

- *ať, (byť), nebť, neboť, nechť, (toť), (vždyť)*

Emphatic -*ž*

Narrower and less fuzzy function than -*ť* – analyzed differently!

- No splitting, kept as one word
- Lemmatized without -*ž*
- New feature: **Emph=Yes**
- But not for lexicalized cases

Examples:

- *dřevniehož, rcěmež, buďž, mějž, přězřiž...*

Except for lexicalized cases:

- *aniž, což, jakož, jehož, kdož, když, kterýž, nikdož, protož, takž, všelicož, ...*

Tokenization

- Opposite situation: Old spelling with space.
  - *kdo koli* “anybody”
  - prepositional phrases vs. compound adverbs

Lemmatization

- Not only picking base form in the paradigm...
- ... but also normalization among alternatives...
- ... even after modernizing orthography!
  - *Křtitel* “Baptist”
  - *Křstitel*
  - *Krstitel*

Lemmatization

- Not only picking base form in the paradigm...
- ... but also normalization among alternatives...
- ... even after modernizing orthography!
  - *Křtitel* "Baptist"
  - *Křstitel*
  - *Krstitel*

- Modern lemma vs. old lemma:
  - forms = *otsúdí* / *otsúdie* "she / they will condemn"
  - lemma candidates = *odsoudit*, *odsouditi*, *odsúditi*, *odsúziti*, *otsoudici*, *otsoudit*, *otsouditi*, *otsúdici*, *otsúdit*, *otsúditi*, *vodsoudit*, *vodsouditi*, *vodsúditi*, *votsoudici*, *votsoudit*, *votsouditi*, *votsúdici*, *votsúdit*, *votsúditi*, ...
  - (modern) lemma = *odsoudit*

Modern Lemma Not Always Easy to Identify

- *jmě* "name" → *jméno*
- *ješutenstvie* "conceit" → *ješitenství*
- *podvstati* → *podvstat?*
- *uvrci* "cast" → *uvrhnout* ... change of inflection class
- *přijieti* "accept" → *přijmout*
- *cizý* "foreign" → *cizí*

- Homonyms with similar meaning:
  - *břich* vs. *břicho* "belly"
  - *obecný* "common" vs. *obecní* "of community"

Negation

Modern Czech data (PDT) until UD 2.14:

- **VERB** productive: *dělat* “to do” → *nedělat* “not to do”
  - Lemma = affirmative form, features: **Polarity=Neg**

Negation

Modern Czech data (PDT) until UD 2.14:

- **VERB** productive: *dělat* "to do" → *nedělat* "not to do"
  - Lemma = affirmative form, features: **Polarity=Neg**
- **ADJ** productive (but competing with antonyms):
  hezký "pretty" → *nehezký* "unpretty, ugly"
  - Lemma = affirmative form, features: **Polarity=Neg**

Negation

Modern Czech data (PDT) until UD 2.14:

- **VERB** productive: *dělat* "to do" → *nedělat* "not to do"
  - Lemma = affirmative form, features: **Polarity=Neg**
- **ADJ** productive (but competing with antonyms):
  hezký "pretty" → *nehezký* "unpretty, ugly"
  - Lemma = affirmative form, features: **Polarity=Neg**
- **ADV**, **DET** somewhat productive:
  nedávno "recently", *nezávisle* "independently"; *neméně* "no less"

Negation

Modern Czech data (PDT) until UD 2.14:

- **VERB** productive: *dělat* "to do" → *nedělat* "not to do"
  - Lemma = affirmative form, features: **Polarity=Neg**
- **ADJ** productive (but competing with antonyms):
  hezký "pretty" → *nehezký* "unpretty, ugly"
  - Lemma = affirmative form, features: **Polarity=Neg**
- **ADV**, **DET** somewhat productive:
  nedávno "recently", *nezávisle* "independently"; *neméně* "no less"
- **NOUN** theoretically all are negatable (even **PROPN**!) but it is rare
  - Meaning is often shifted
  - Annotation not completely consistent
    - **Inflection:** *nezávislost* "independence", *neschopnost* "incompetence", *nemožnost* "impossibility", *nedodržení* "nonobservance", *neznalost* "unawareness"
    - **Derivation:** *nebezpečí* "danger", *nedostatek* "shortage", *nezaměstnanost* "unemployment", *nemovitost* "real estate", *nesmysl* "nonsense", *neštěstí* "misfortune, accident"

Negation

Old Czech data and Modern Czech data (PDT) since UD 2.15:

- **NOUN** does not have **Polarity**, lemma with *ne-* ⇒ negation is derivation

- All non-pronominal **ADV** have **Polarity** (*nevelmi* "not much" – this can be also quantifier, i.e., **DET**)
- Some pronominal **ADV** have it, too (*nevždy* "not always")

Morphological Features

Simple Past Tense: Imperfect

- Old: *Ale Kristovo porozenie tak bieše.*
  - **VERB** VerbForm=Fin Mood=Ind Tense=Imp Aspect=Imp Voice=Act Number=Sing Person=3 Polarity=Pos
- Modern: *S narozením Ježíše Krista to bylo takto:*
  - **VERB** VerbForm=Part Tense=Past Aspect=Imp Voice=Act Number=Sing Gender=Neut Polarity=Pos
- English: "Now the birth of Jesus Christ took place in this way."

Morphological Features

Simple Past Tense: Imperfect

- Old: *Ale Kristovo porozenie tak bieše.*
  - **VERB** VerbForm=Fin Mood=Ind Tense=Imp Aspect=Imp Voice=Act Number=Sing Person=3 Polarity=Pos
- Modern: *S narozením Ježíše Krista to bylo takto:*
  - **VERB** VerbForm=Part Tense=Past Aspect=Imp Voice=Act Number=Sing Gender=Neut Polarity=Pos
- English: "Now the birth of Jesus Christ took place in this way."

Simple Past Tense: Aorist

- Old: *Tehdy oni povědĕchu jemu:*
  - **VERB** VerbForm=Fin Mood=Ind Tense=Past Aspect=Perf Voice=Act Number=Plur Person=3 Polarity=Pos Variant=Long
- Modern: *Oni mu řekli:*
  - **VERB** VerbForm=Part Tense=Past Aspect=Perf Voice=Act Number=Plur Gender=Masc Animacy=Anim Polarity=Pos
- English: "They told him,"

Conditional vs. Aorist of “to be”

- **Aorist** of *být* “be”

	Sing	Dual	Plur
1	*bych*	*bychově*	*bychom*
2	*by*	*bysta*	*byste*
3	*by*	*bysta*	*bychu*

*) and other variants

- Modern **Conditional** auxiliary

	Sing	Plur
1	*bych*	*bychom*
2	*bys*	*byste*
3	*by*	*by*

- Even modern *by* can be 2nd person in *by ses*
- In Old Czech, aorist servers as conditional auxiliary...
  - ... but *by* is used in various persons and numbers
  - Aorist usages must be distinguished from conditional

Morphological Features

Dual Number

- Old: *... uzřě dva bratry, ... že biešta rybářě.*
  - **NOUN Gender=Masc Animacy=Anim Number=Dual Case=Nom**
- Modern: *... uviděl dva bratry, ... byli totiž rybáři.*
  - **NOUN Gender=Masc Animacy=Anim Number=Plur Case=Nom**
- English: "... he saw two brothers, ... for they were fishermen."

Morphological Features

Dual Number

- Old: *... uzřě dva bratry, ... že biešta rybářě.*
  - **NOUN Gender=Masc Animacy=Anim Number=Dual Case=Nom**
- Modern: *... uviděl dva bratry, ... byli totiž rybáři.*
  - **NOUN Gender=Masc Animacy=Anim Number=Plur Case=Nom**
- English: "... he saw two brothers, ... for they were fishermen."

Animacy

- Not significant grammatically as in modern Czech
- But tentatively annotated anyway, to be consistent with modern Czech data

Morphological Features

Converbs (= Gerunds = Transgressives)

- Old: *Tehdy ona ihned ostavše sieti, jidesta po něm.*
  - **VERB** VerbForm=Conv Tense=Past Aspect=Perf Voice=Act Number=Dual Polarity=Pos
- Modern: *Oni hned opustili sítě a následovali ho.*
  - **VERB** VerbForm=Part Tense=Past Aspect=Perf Voice=Act Number=Plur Gender=Masc Animacy=Anim Polarity=Pos
- English: "Immediately they left their nets and followed him."

Morphological Features

Converbs (= Gerunds = Transgressives)

- Old: *Tehdy ona ihned ostavše sieti, jidesta po něm.*
  - **VERB** VerbForm=Conv Tense=Past Aspect=Perf Voice=Act Number=Dual Polarity=Pos
- Modern: *Oni hned opustili sítě a následovali ho.*
  - **VERB** VerbForm=Part Tense=Past Aspect=Perf Voice=Act Number=Plur Gender=Masc Animacy=Anim Polarity=Pos
- English: "Immediately they left their nets and followed him."

Accusative Converbs?

- Old: *... někteří ... neuzřie syna člověčieho, přijdúce v svém království.*
  - **VERB** VerbForm=Conv Tense=Pres Aspect=Perf Voice=Act Number=Sing Gender=Masc Animacy=Anim Polarity=Pos ~~Case=Acc~~
- Modern: *... někteří ... nespatří Syna člověka přicházejícího ve své královské moci.*
  - **ADJ** VerbForm=Part Tense=Pres Aspect=Imp Voice=Act Number=Sing Gender=Masc Animacy=Anim Polarity=Pos Case=Acc
- English: "... some ... (will not) see the Son of Man coming in his kingdom."

- Decision: No Case feature with converbs.

Morphological Features

Supine vs. Infinitive (???)

- Old: *Nalezeny sú oslice, jíchžtos byl šel hledat.*
  - **VERB** **VerbForm=Sup Aspect=Imp Polarity=Pos**
- Modern: *Oslice, které jsi šel hledat, se našly.*
  - **VERB** **VerbForm=Inf Aspect=Imp Polarity=Pos**
- English: "The donkeys that you went to seek are found."
- UD knows supine in Modern Slovenian.
- But what criteria would we use in Czech?

Summary

- 100K UD-style etalons to train UDPipe for each period
- Current work:
  - The ÚJČ team + DZ: Old Czech
  - The ÚČNK team + BŠ: 19th century
  - The NK team: preparing data from 20th century
- Modern lemma to facilitate searching
- Morphological features for extinct phenomena
- Tokenization: Multi-word tokens

https://korpus.cz/hickok