

# DIGITIZATION PRACTICE IN THE CZECH NATIONAL CORPUS

Anna Řehořková

[anna.rehorkova@korpus.cz](mailto:anna.rehorkova@korpus.cz), Institute of the Czech National Corpus

*Optical Character Recognition and Text Encoding for ELTeC*

*Contributions Training School*

2018-04-16





1994...



*Český národní korpus*

KonText | SyD | Morfio | KWords | Treq | Wiki | Support | Biblio

Anna Řehořková | Logout | český

☒ ignore case
 ☐ search word forms of the given lemma

### SYN RELEASE 6

DECEMBER 18, 2017

Release 6 of the corpus of contemporary written Czech [SYN](#) has been published on 18th December 2017. As a result of inclusion of journalistic texts from 2016, size of the SYN [release 6](#) has exceeded 4 billion words.

### KONTEXT: VERSION 0.11

DECEMBER 15, 2017

An [updated version](#) of the KonText interface is now available with new functions, such as 2-dimensional frequency distribution with confidence intervals or improved query history where individual items can be archived under a custom name.

### SLAVICORP 2018 CONFERENCE

DECEMBER 6, 2017

Institute of the Czech National Corpus is proud to announce that the SlaviCorp conference will take place in Prague on September 2018. Call for papers and all other information can be found at the [conference web](#).

## What is a corpus?

A language corpus is an electronic collection of authentic texts (written or spoken) easily searchable for various language phenomena (esp. words and collocations) and to display them in their natural context.

The [CNC corpora](#) include written contemporary Czech (more than 4 billion tokens), spontaneous spoken language (more than 7 million tokens), diachronic corpus of historical texts and parallel corpus InterCorp that contains translations from or to 30+ languages.

[more...](#)

## Who are we?

The Czech National Corpus is an **academic project** founded in 1994 at the [CU FA](#) and administered by the [Institute of the Czech National Corpus](#). The aim of the project is systematic mapping of Czech and other languages in comparison with Czech. CNC corpora are accessible to everybody interested in studying the language after [free registration](#).

[more...](#)

## Support and information resources

# Available corpora

[My list](#) | [All corpora](#) X

Reset

SYN series

ORAL series

InterCorp

synchronic

diachronic

spoken

written

web

current version

older version

Czech

non-Czech

parallel

comparable

author

representative

learner

specialized

Hold CTRL/Command to select multiple labels

dia

- diakon 145M
- diakorp v6 4M
- diakorp v5 2M
- dialekt v1 - dial 128k
- dialekt v1 - ort 126k
- dialog 1M

Hit [Tab] to see your favourite corpora



# https://kontext.korpus.cz



KonText

SyD

Morfio

KWords

Treq

| Wiki

Support

Biblio

Anna Řehořková

Logout | český



Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: syn2015 | Query: XML (94 hits)


Hits: 94 | i.p.m. 0: 0.78 (related to the whole "syn2015") | ARF 0: 14.86 | Result is sorted

1 / 2

Line selection: simple | Attributes:

<input type="checkbox"/>		NFC: oborová literatura	může být u různých CRM systémů realizováno buď přes standardní	XML	rozhraní , či prostřednictvím datového skladu . Například pro firmy
<input type="checkbox"/>		NFC: oborová literatura	je vytvořen v modulu Účetnictví , který též , prostřednictvím	XML	, komunikuje s finančním úřadem a vloží údaje do excelu
<input type="checkbox"/>		NFC: oborová literatura	doručovaná tímto způsobem musí být ve zveřejněném formátu a struktuře	XML	souboru ( viz Pokyn č. D - 349 ) .
<input type="checkbox"/>		NFC: oborová literatura	Pokyn č. D - 349 ) . Pro vytvoření požadovaného	XML	souboru lze využít elektronické formuláře aplikace EPO . </p><p> RUČENÍ ZA
<input type="checkbox"/>		NFC: oborová literatura	Příkladem je segment metadat , zapsaných ve formátu značkovacího jazyka	XML	. Metadata , vytvořená v prostředí ArcGIS , popisují přítomnost
<input type="checkbox"/>		NFC: oborová literatura	, je vztahem mezi jedincem a datovým typem definovaným podle	XML	Schematu ( Smith 2004 ; Biron 2004 ) nebo vztahem
<input type="checkbox"/>		NFC: oborová literatura	jsou na principu Request / Response . </p><p> obvykle využívají jazyk	XML	( SOAP ) , ale mohou využívat i jiný způsob
<input type="checkbox"/>		NFC: oborová literatura	standardem , protokolem ) . Celou architekturou se prolíná jazyk	XML	, který je používán pro definování většiny implementačních protokolů a
<input type="checkbox"/>		NFC: oborová literatura	. </p><p> Čtvrtá metoda je již zmíněná metoda , která vrací	XML	dokument s popisem služby založeným na odpovědi dle OGC WMS
<input type="checkbox"/>		NFC: oborová literatura	je využíván protokol SOAP v režimu Request-Response . Následující ukázky	XML	kódů představují Request a Response takového volání na náhodně zvolené
<input type="checkbox"/>		NFC: oborová literatura	pro rychlou orientaci bychom v tomto případě pár zařadili jako	XML	) . </p><p> • Jestliže u posuzované kolekce chybí některý charakteristický
<input type="checkbox"/>		NFC: oborová literatura	v prostředí Autodesk Inventoru . Exportním příkazem Inventor systém vygeneruje	XML	soubor se všemi kabelem , Pro malé kanceláře a domácnosti
<input type="checkbox"/>		NFC: oborová literatura	finální tvorba titulků v rozlišení 2K . Export metadat do	XML	formátů . Export soupisek a textových výstupů pro kontrolu skriptů
<input type="checkbox"/>		NFC: oborová literatura	na synchronním , tak na asynchronním předávání zpráv ve formátu	XML	při zajištění plné bezpečnosti pomocí kvalifikovaných certifikátů . * </p><p> Názorný
<input type="checkbox"/>		NFC: oborová literatura	a analýzu dat nacházejících se v relačních , OLAP a	XML	zdrojích dat . Obohacuje komfort koncového uživatele o novou integrovanou
<input type="checkbox"/>		NFC: oborová literatura	všechny běžné formáty kancelářských klasických aplikací , dále HTML ,	XML	, SGML , textové formáty pro PC i Mac ,
<input type="checkbox"/>		NFC: oborová literatura	to všechny běžné formáty kancelářských aplikací , dále HTML ,	XML	, SGML , textové formáty pro PC i Mac ,
<input type="checkbox"/>		NFC: oborová literatura	, které specifikují konkrétní reprezentaci přenášených znalostí , jako je	XML	( eXtensible Markup Language ) , KIF ( Knowledge Interchange
<input type="checkbox"/>		NFC: oborová literatura	o seznam senzorů , z nichž každý má svůj vlastní	XML	konfigurační soubor se seznamem konkrétních datových skladů a parametry potřebnými
<input type="checkbox"/>		NFC: oborová literatura	silnějším stránkám FlexiBee . </p><p> Veškeré sestavy lze uložit ve formátu	XML	a PDF , faktury lze navíc exportovat do formátu ISDOC
<input type="checkbox"/>		NFC: oborová literatura	pro elektronické podání ve formátu definovaném ministerstvem financí . Vygenerovaný	XML	soubor lze následně odeslat na portál České daňové správy nebo
<input type="checkbox"/>		NFC: oborová literatura	v Byznysu je také zapracování exportu účetních výkazů do souboru	XML	ve struktuře pro Centrální systém účetních informací státu . </p><p> IFS
<input type="checkbox"/>		NFC: oborová literatura	a analýzu dat nacházejících se v relačních , OLAP a	XML	zdrojích dat . Zdokonalené uživatelské rozhraní nově nabízí komplexní řadu

# https://syd.korpus.cz

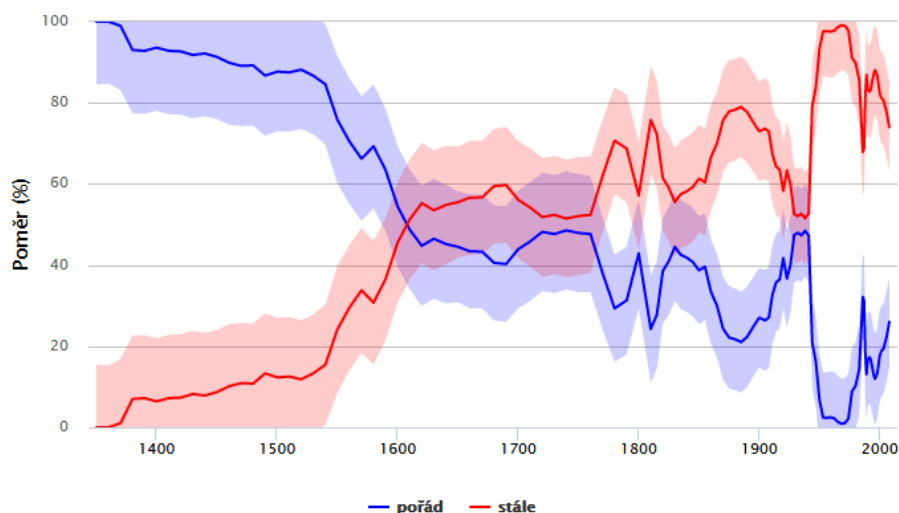
SyD 

[Diachronní](#) [Synchronní](#) [Nový dotaz](#)

[1] pořád [2] stále +

Okno: proměnlivé Krok: proměnlivý

[Porovnat varianty](#) ☒ a=A ☐ lemma [Zobrazit adresu dotazu](#)



Zkoumání proměn poměru variant v čase je založeno na korpusu *Diakon*, který je sestaven na základě textů korpusu SYN a diachronní složky ČNK tak, aby v rámci současných možností co nejlépe pokrýval celé období existence psané češtiny. Uživatele je třeba upozornit na skutečnost, že texty diachronní složky ČNK zařazené do korpusu *Diakon* z velké části nebyly dosud zkorigovány, a výsledky vyhledávání konkurence variant nelze proto považovat za zcela přesné, i když už v současné době umožňují získat vysoce spolehlivou představu o celkovém vývoji. *Diakon* zahrnuje texty od nejstarších památek ze 14. století (nejstarším je *Alexandreida*, datovaná do roku 1310) až po publicistiku z let 1991-2009. Vzhledem k proměnám písemné produkce, kterými čeština za 7 století své existence prošla, není možné klást na takový korpus stejné požadavky stylové a žánrové vyváženosti, jaké máme na korpusy češtiny současné.

Úvodní graf prezentuje tendence vývoje poměru variant v průběhu času. Každá linie reprezentující jednu variantu je obklopena stejnobarevnou oblastí, jejíž šíře značí spolehlivost naměřených dat (čím širší oblast, tím větší možná chybovost výsledku).

[víc](#)

Kluzavý průměr: 1  ☒ zobrazit chybu [Uložit jako](#) ▼

[Celá historie](#) [14.-18. stol.](#) [19. stol.](#) [1900-1989](#) [1990-2009](#)

[Souhrn](#) [Pokrytí](#) [FQ / roky](#) [Tabulka](#)

[1] frekvence	14002
pořád	12227
Pořád	1760
POŘÁD	15

[2] frekvence	58827
stále	54407
Stále	4310
STÁLE	110

Absolutní frekvence všech slovních tvarů odpovídajících jednotlivým dotazům v korpusu *Diakon*.

# https:// **IKI** words.korpus.cz

**Language of target text:** English ▾ Help

**Target text (the text you want to examine):**  
▶ Enter text  
▶ Upload text file(s) / Multi-analysis

**Reference corpus (background to which the text is compared):**  
▾ Select reference corpus  
BNC ▾  
▶ Enter reference text  
▶ Upload reference text file

**Stop-list (words excluded from the analysis):**  
☐ pronouns  
☒ prepositions  
☒ conjunctions  
☒ numbers  
☒ ignore case  
Analyze  
Settings

Text Keywords Distribution Keyword links Concordance

## Corpus selection

The selection of a **corpus** suitable for solving the given research question is an important decision which must be made before taking any other steps in the research. The range of **corpora** made available through the project of the Czech National **Corpus** is constantly widening, as can be seen on the list of **corpora**. It has therefore been necessary to adjust the **corpus selection** on the KonText interface in order to accommodate their growing number. Until the autumn of 2015, the **corpus selection** had the shape of a hierarchically organized tree; this system had several disadvantages: from the not always definite placement of the given **corpus** within the hierarchy, to a great increase in the number of new **corpora** and their versions. As a result of this, the hierarchical organization stopped being clear and sustainable in the future, which is why we have switched to the new, label-based system. Its aim is primarily to facilitate orientation in the large number of existing **corpora**, while simultaneously simplifying work for those users who use only a small number of favorite **corpora**.

After clicking on the name of the **corpus** (in the default setting it is always the most recent representative **corpus** of synchronic written Czech, currently SYN2010) a frame for the selection of a **corpus** appears, containing two main sections:

### Corpus selection: featured and favorite corpora

My list with a quick, single click selection of **corpora**. This quick selection contains favourite **corpora**, which can be selected by the user, and also the so-called featured **corpora**: a default list of several **corpora**, which the CNC considers to be particularly important in the individual areas of production. Having them all in one place simplifies the selection of a **corpus** especially for beginning users of the CNC. Favourite **corpora** can be selected either on the page with all the available **corpora**, or when working with them at the time of query input (such **corpora** are labelled with a yellow star).

All **corpora** with the possibility of searching all available **corpora** with the aid of so-called labels, which are used to characterize the **corpora** (a typical **corpus** has several labels, e.g. SYN2010: written, synchronic, Czech, SYN family, representative). For example, if you are looking for a web **corpus** of Czech, all you have to do is select the labels "Czech" + "web", and all the relevant **corpora** made available by the CNC will appear. The search may be further refined by typing part of the **corpus name** or its description into the search bar, and the resulting list of **corpora** is interactively filtered based on the keywords. However, it must be noted that for spatial reasons the list only shows the first 25 items; if the list is too long, the query must be specified further with the addition of another label, or by searching for a part of its name.

Example: The user searches in the tab All **corpora** for a current version of the English section of the parallel **corpus** InterCorp. He first selects the labels "InterCorp" a "current version", the first 25 **corpora** which conform to the specified conditions appear on the list, although InterCorp contains many more languages. The **corpora** not displayed can be accessed with the help of further filtering, for example by typing part of the **corpus name** or language (please note that the names of the individual InterCorp versions are in English). After finding the desired **corpus** and clicking on it, the **corpus** becomes the current **corpus** for searching, and it is at the same time possible to mark it as favourite with a star. The **corpus** is added to the list of favourite **corpora** and can be accessed quickly and easily with a single click.

# Made by ICNC

## CZECH

- synchronic written
- synchronic spoken
- diachronic written
- dialect spoken

## OTHER LANGUAGES

- parallel





# Made by ICNC

## CZECH

- synchronic written
- synchronic spoken
- diachronic written
- dialect spoken

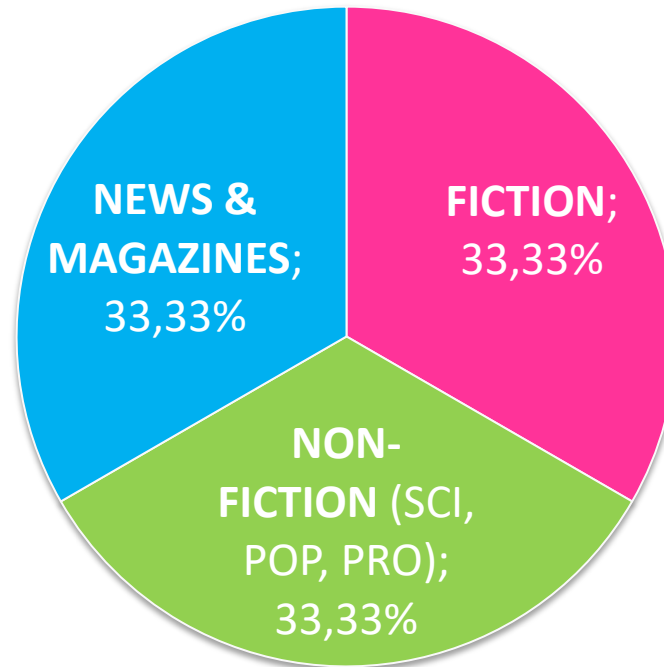
## OTHER LANGUAGES

- parallel



# Contemporary language: SYN series

- SYN2000, SYN2005, SYN2010, SYN2015
- 100 mil. words



# Contemporary language: text processing

- .doc, .docx, .rtf, .txt, .pdf, .epub...

Typical example:

- 1) machine-readable format with markup →  
conversion to HTML via LibreOffice
- 2) HTML → TEI intermediate format



# HTML

<P CLASS...

<SPAN STYLE="text-transform: uppercase"><FONT FACE="Times New Roman, serif"><FONT SIZE=4><B>První část</B></FONT></FONT></SPAN></P>

<P CLASS="western" STYLE="margin-bottom: 0in; line-height: 150%"><BR>  
</P>

<P CLASS="western" ALIGN=CENTER STYLE="margin-bottom: 0in; line-height: 150%">  
<FONT FACE="Times New Roman, serif"><FONT SIZE=4>Kapitola 1</FONT></FONT></P>



# TEI Intermediate format

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<doc xmlns="http://www.korpus.cz/imfSchema"  
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"  
  xsi:schemaLocation="cnkImfSchema.xsd">
```

```
<text type="body">
```

```
<p><pb n="103"/></p>
```

```
<p><hi rend="bold">První část</hi></p>
```

```
<p>Kapitola 1</p>
```

```
<p>Hrad se rozpadal, ale ve dvě ráno to v nanicovatém světle měsíce Danny  
neviděl. To, co měl před sebou, vypadalo pekelně solidně: dvě kulaté věže,  
mezi nimi oblouk, a ten uzavírala železná brána, která se možná nepohnula  
posledních tři sta let, nebo nikdy.</p>
```





# TEI Intermediate format → output format

<p type=normal id=egano\_hrad:1:1><hi rend=bold>  
První část </hi></p>

<p type=normal id=egano\_hrad:1:2> Kapitola 1 </p>



# Contemporary language: text processing

- .doc, .docx, .rtf, .txt, .pdf, .epub...

Typical example:

- 1) machine-readable format with markup → conversion to HTML via LibreOffice
- 2) HTML → TEI intermediate format
- 3) series of scripts to handle specific errors



# Wrongly divided words

<p>Vyrážím. Můj tmavý odraz mě pronásleduje až do rohu výkladu a mizí. V jednom titulku čtu: <hi rend="italic">„ Takto budete letos **vy-padat**."</hi> Cigareta na mě působí báječně a já se pomalu znovu stávám člověkem. Cítím se cize a čerstvě, jako vůně nových bankovek, osvobozený od všeho, co bývalo. Pohyby a myšlenky se rozplývají.</p><p>Tohle je začátek.</p><p>Autobusová zastávka u náměstí Gullmarsplan, nechci jet metrem. Nejsem venku, na svobodě, ještě ani dvě hodiny a už tu stojím jako kdokoliv jiný, a čekám na autobus. Téměř **usí-nám**. Přechody mezi euforií a nepopsatelnou únavou jsou nebezpečné, toho jsem si vědom. Nejsem oholený a ani **ne-chci** myslet na to, jak vypadají moje vlasy. Podrážka umrtvuje cigaretu.</p>



# Concentration of non-alphabetic signs in a paragraph

<p>Šel spát s rozviklaným sebevědomím.</p>

<p><hi rend="italic">\*</hi> \* \*</p>

<p>Dá se v těchto krátce načrtnutých situacích  
mluvit o střídání pesimismu s optimismem?



# Foreign-language filter

Do druhé kategorie je naopak možno zařadit zpravidla mimo země bývalého Sovětského svazu, o kterých je řeč, vzniklé dokumentární práce, jako jsou monumentální úvod do postsovětského prostoru od nestorů americké sovětologie nebo zvláštní příloha dokumentů k sborníku o místu a roli krymské otázky v rusko-ukrajinských vztazích.

<note place="foot"><p><hi rend="italic">Russia and the Commonwealth of </hi><hi rend="italic">Independent States. </hi><hi rend="italic">Documents, Data, and Analysis,</hi> ed. Zbigniew Brzezinski, Paige Sullivan (Armonk, NY: M.E. Sharpe, The Center for Strategic and International Studies, 1997); <hi rend="italic">Crimea: Dynamics, Challenges and Prospects,</hi> ed. Maria Drohobycky (Lan- ham, MD: Rowman and Littlefield, 1995).</p></note></p>





# Historical languages: DIA group

- **unstructured archives:**

1. diakorp

- manually corrected, tagging of text structure (headlines, foot notes etc.)

2. diakon (SyD)

- only automatic correction of non-Czech signs

- **DIA series (in prep.)**

- ballanced for text types (FIC, NFC, NMG)
- 19<sup>th</sup> century, core: 1 mil. words



# Historical language: text processing

- scans (.pdf, .png, .jpg, .djvu...)

Typical example:

- scans → ABBYY FineReader OCR → manual check against the scan



tedy tomu ubrániti náš jezdec, sedě mimo to na dobrém koni?

Avšak nutno, abychom se s ním seznámili. Jezdec byl mladý šlechtic od heidelbergského dvora, jménem svobodný pán z Elvangu, důvěrník to kurfiřta falckého. Všude v Německu panovalo mínění, že evangelické obyvatelstvo v Čechách neuspokojí se pouze ohrazením proti volbě Ferdinanda za budoucího nástupce v království. Největší napnutí stran této záležitosti panovalo však u heidelbergského dvora, a ačkoliv kurfiřtovi radové s nejčelnějšími českými protestanty si dopisovali, přec uznal Bedřich zahodno, poslati do Čech pozorovatele, aby co možná nejurčitěji o tamnějším smýšlení a ruchu se poučil. K tomu určil právě tohoto mladého šlechtice, s nímž v mládí již důvěrně byl obcoval. Kurfiřt zvolil jej k tomu cíli, poněvadž jemu co nejhorlivější přivrženec evangelického učení u heidelbergského dvora znám byl. Reformace měla pro mladého šlechtice veliké důležitosti proto, poněvadž skrz ní bylo vyrčeno vyproštění z jařma římského, které Německo a českou zemi po

mnoha století v duševním ohledu potlačovalo a v tělesném ohledu co nejhnusněji olupovalo, a poněvadž nové učení zavržením učení o neomylnosti papeže a podáním bible jakožto knihy pro každého z lidu vyslovilo právo, že možno k spasiteli i bez zprostředkování kněžské vlády se přiblížiti a slovo jeho na sebe nechat účinkovati. Domníval se, že právě tak jako zářící slunce na kopcích, v údolí a planině vše k pučení a k životu vzbuzuje, i duch písma svatého, nenahromadí-li se mlhy kněžské nadutosti a kněžské vlády mezi ním a lidem, vzbudí bohatý duševní život. A jak by se bylo co obávati života vyvíjejícího se bezprostředním účinkováním ducha božího, vanoucím ve slově biblickém? Třebas by omylové tu neb onde jako koukol povstali, přec by požehnání muselo tisíckrát násobně převažovati. Mnozí však předstírají takové obávání pouze proto, poněvadž se strachují hrozící ztráty nepravě nabytých světských výhod. V tom záležela dle mínění šlechtice hlavní příčina ouz-kostného volání, zaznívajícího z Říma.

Takovými myšlenkami byl i dnes Elvangu opanován při pohledu na čerstvý.

Anonymous: *Bitva na Bílé hoře* (The Battle of White Mountain), 1864



šlechtice, e nímž v mládí již důvěrně byl obcoval. Kurfiršt zvolil jej k tomu cíli, poněvadž jemu co nejhorlivější přívrženec evangelického učení u heidelbergského dvora znám byl. Reformace měla pro mladého šlechtice veliké důležitosti proto, poněvadž skrz ní bylo vyrieno vyproštění z jařma římského. které Německo a českou zemi po mnoha století v duševním ohledu potlačovalo a vfelesném ohledu co nejhnusněji olupovalo, a poněvadž nové učení zavržením učení o neomylnosti papeže a podáním bible jakožto knihy pro každého z lidu vyslovilo právo, že možno k spasiteli i bez zprostředkování kněžské vlády) se přiblížiti a slovo jeho na sebe nechat účinkovati. Domníval se, že právě tak jako zářící slunce na kopcích, v údolích a planině vše k pučení a k životu vzbuzuje, i duch písma svatého, nenahromadí-li se mlh) kněžské nadutosti a kněžské vlády mezi ním a lidem, vzbudí bohatý duševní život. A jak by se bylo co obávati života vyvinujícího se bezprostředním účinkováním ducha božího, vanoucím ve slově biblickém? Třebas by omylové tu neb onde jako koukol povstali, přec by požehnání muselo tisíckrát násobně převažovati. Mnozí však předstírají takové obávání pouze proto, poněvadž se strachují

Anonymous: *Bitva na Bílé hoře* (The Battle of White Mountain), 1864



# Historical language: OCR output

šlechtice, **e→s** nímž v mládí již důvěrně byl obcoval. Kurfiršt zvolil jej k tomu cíli, poněvadž jemu co nejhorlivější přívrženec evangelic-kého učení u heidlsbergského dvora znám byl. Reformace měla pro mladého šlech-tice veliké důležitosti proto, poněvadž skrz **ní→ni** bylo **vyrieno→vyrčeno** vyproštění z jařma řím-ského. které Německo a českou zemi po mnoha století v duševním ohledu potlačovalo a **vfelesném→v tělesném** ohledu co nejhnusněji olupovalo, a poněvadž nové učení zavržením **učení→učení** o neo-mylnosti papeže a podáním bible jakožto knihy pro každého z lidu vyslovalo právo, že možno k spasiteli i bez zprostředkování kněžské **vlád)→vlády** se přiblížiti a slovo jeho na sebe nechat účín- kovati. Domníval se, že právě tak jako zářící slunce na kopcích, v údolí a planině vše k pučení a k životu vzbuzuje, i duch písma svatého , nenahromadí-li se **mlh)→ mlhy** kněžské nadutosti a kněžské vlády mezi **nim→ním** a lidem, vzbudí bohatý duševní život. A jak by se bylo co obávati života vyvi-nujícího se bezprostředním účinkováním ducha božího, vanoucím ve slově **biblic- %émT→biblickém?** Třebas by omylové tu neb onde jako koukol povstali, přec by požehnání muselo tisíckrát násobně převažovati. Mnozí však předstírají takové obávání pouze proto, poněvadž se **strachuji→strachují**





## ***Majka obecná I. 4.***

Toho brauka mnohý z wás giž znáti bude, zwlá-  
ště když wás upamatugi, že řjkáwáte, když gste ho  
někde chytili: „Magko, magko,

Dey mi masti  
Na bolesti,“

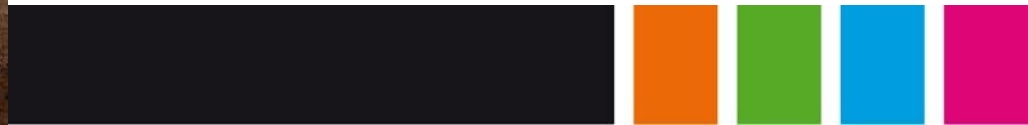
což tak dlaugo opakujete, a gj postrkowánjm nutjte,  
až gj všude z klaubečků, totiž zvláště z koljnek a  
pat atd. co wjno žluté kapičky se wypocugj. A o-

<k>Majka obecná 1. <e>I.</e> 4.</k>

Toho brouka mnohý z vás již znát bude, zvláště když vás  
upamatuji, že říkáváte, když jste ho někde chytili: <v>„Majko, majko, dej  
mi masti na bolesti,“</v> což tak dlouho opakujete a ji <e>gj</e>  
postrkováním nutíte, až jí všude z kloubečků, totiž zvláště z kolínek a pat  
a t. d., co víno žluté kapičky se vypocují.



- no reliable dictionary for the old language (neither electronic nor a paper one)
- insufficient quality of print
- ICNC grant (partially) → text transcription
- building a dictionary bottom-up from all correct word forms in a text



# Thank you!

