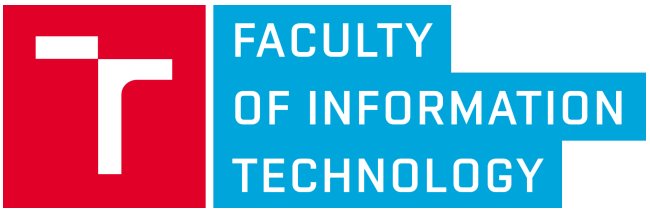


Technická Dokumentace k Software TextBite – Systém pro analýzu struktury dokumentů

Martin Kostelník, Karel Beneš, Michal Hradiš, Marek Vaško

Vysoké učení technické v Brně Brno
2023

Tento dokument byl vytvořen s finanční podporou MK ČR v rámci programu **NAKI III program na podporu aplikovaného výzkumu v oblasti národní a kulturní identity na léta 2023 až 2030**

v projektu semANT - Sémantický průzkumník textového kulturního dědictví.

Číslo a název projektu:

DH23P03OVV060	semANT - Sémantický průzkumník textového kulturního dědictví

Název a popis dílčího výstupu:

TextBite – Systém pro analýzu struktury dokumentů
Tento dokument popisuje funkčnost a použití software TextBite, který extrahuje sémanticky souvislé části dokumentů.

Jazyk dokumentu

Angličtina

Organizace a řešitel

Vysoké učení technické v Brně	Ing. Michal Hradiš Ph.D.

Availability

The software module is available from https://github.com/DCGM/semANT-TextBite.

Python distribution package is available for download at https://pypi.org/project/textbite, installed as "`pip install textbite`".

License

BSD 3-Clause License

Usage

This software provides a semantic layout analysis on top of plain OCR output. TextBite enhances a PAGE XML description of an analyzed page by introducing title elements, clustering text lines in semantically related parts (chapters, articles, dictionary entries, …), reading order and altering already present regions as needed. All of this new information is stored in a standard way described by the PAGE standard, allowing for further processing. See Fig. 1 for an overview of the process.

pzpi

Fig 1. Application of TextBite on a new input document. First, the Pero OCR is used to get the basic layout of the document, TextBite is then applied to extract logical elements.

The package is designed for usage as a command line application, processing pages in batches, as given by folders of files.

Requirements

The core functionality is platform-independent.

Python in version 3.7 or higher is required, alongside packages `pero-ocr` and `ultralytics` on top of common packages such as `numpy`. All of these requirements are covered by the standard installation procedure.

There is no compute-heave operation in TextBite that would require the usage of acceleration hardware, any modern CPU is sufficient for reasonable speed of operation.

Technical solution

The core of TextBite is a detector model based on YOLOv8. This detector identifies logical chunks directly in the image of the page. These detections are then merged with the available region and textline information to provide an enhanced page representation.

To train the detector, we have collected a custom dataset of publicly available pages from the Czech Digital Library. To promote diversity of the dataset, pages were specifically sampled to cover periodicals, dictionaries and books as major classes of documents, complemented by completely random pages from the whole collection. These pages were annotated for logical units by volunteers. For the current version of TextBite, ca. 1600 pages were used for training and ca. 100 were kept for validation and testing each.

Regions identified by the volunteers as logical units were then aligned with textlines detected by the Pero system to provide a precise annotation for training and evaluating the detector model.

When deployed, TextBite operates in five steps: (1) The detector provides rectangular predictions of continuous logical parts. These (2) are aligned with textlines provided in the corresponding XML. As needed, text regions in the XML (3) are refined to match the logical boundaries provided by the detector. Once this is done, we conservatively (4) link the individual parts together, typically merging text regions with their preceding titles. Finally, this information (5) is stored in the XML as described below in section Output format. Breakdown of accuracy of the system is provided in Table 1, several commented examples are given at the end of this document.

Table 1: Performance of TextBite on validation data, reported as V-measure between predicted segmentation and ground-truth [%]. Note that document categories are sorted by number of examples, i.e. there is the least number of book pages – motivated by the fact that there is the least variability in them, whereas newspapers come in very many different sizes and layouts.

Page type	Books	Dictionaries	Periodicals
# pages	15	30	45
V-measure	73.4	88.3	84.6

During development, we have experimented with clustering of textlines and/or text regions based on graph neural networks, however these approaches have not yielded a reliable performance. Enforcing robustness in these models is the next step in enhancing capabilities of TextBite in the future.

Publicly available model

We publish a detection model trained on a diverse mixture of various documents annotated for logical chunks of data (books, periodicals of various layouts, dictionaries). The model is available online. The overall V-measure of TextBite using this particular model is 86 %, measured on validation data.

Software organization

This software has a single entry point – the executable `textbite`. In case TextBite was not pip-installed, explicit paths need to be provided as per `pyproject.toml`.

The invocation of the executable looks as follows:

```
textbite --model model.pt --images-input pages-img/ --xml-input pages-xml/ --xml-output textbite-out/
```

Here, the individual parameters stand for:

model	A file with the detector model. **Optional:** If not given, the public model is automatically downloaded (internet connection needed).
images-input	A folder with images of pages to be analyzed
xml-input	A folder with corresponding PAGE XML results of OCR applied to said pages.
xml-output	A folder for updated PAGE XMLs. In case it is the same as xml-input, the XMLs are overwritten.

Output format

TextBite enhances the information in the PAGE XML in two ways: It (1) labels regions containing article titles, chapter headings etc. using the type field of the region, e.g.: `<TextRegion id="r002" type="heading">` and (2), it introduces an explicit reading order that groups regions constituting the logical chunks, e.g.:

```
<ReadingOrder>
    <UnorderedGroup id="root">
        <OrderedGroup id="bite_1">
            <RegionRefIndexed regionRef="r002" index="0"/>
```

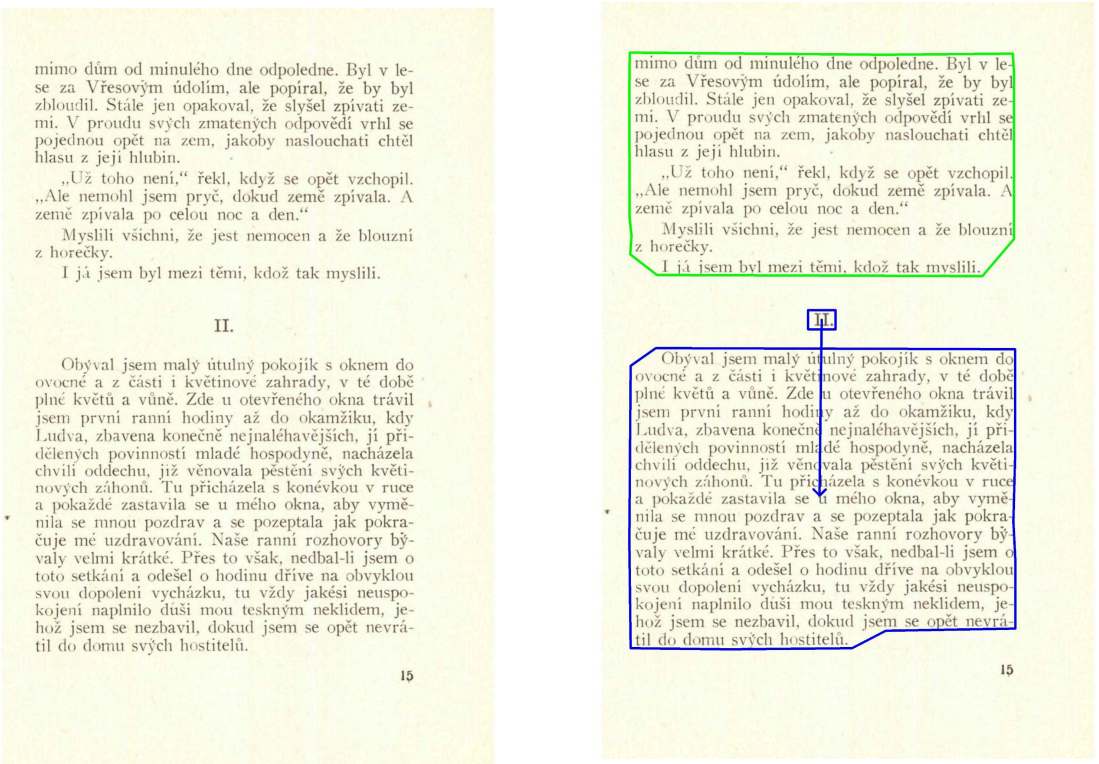
```
                <RegionRefIndexed regionRef="r000"
            index="1"/>

            </OrderedGroup>

            <OrderedGroup id="bite_2">

                <RegionRefIndexed regionRef="r003"
            index="0"/>

                <RegionRefIndexed regionRef="r001"
            index="1"/>

            </OrderedGroup>

      </UnorderedGroup>

</ReadingOrder>
```

Note that some elements of the page, such as a page number or an entries span in a dictionary may be left out as they are not part of any of the semantic units.

These changes are in line with the PAGE XML definition.

Examples

Technically TextBite enhances PAGE XML description of the analyzed page (such as can be obtained from pero-ocr application). Here, we present some examples by encoding this information into coloring of individual regions.



mimo dům od minulého dne odpoledne. Byl v lese za Vřesovým údolím, ale popíral, že by byl zbloudil. Stále jen opakoval, že slyšel zpívati zemi. V proudu svých zmatených odpovědí vrhl se pojednou opět na zem, jakoby naslouchati chtěl hlasu z jeji hlubin.

„Už toho není," řekl, když se opět vzchopil. „Ale nemohl jsem pryč, dokud země zpívala. A země zpívala po celou noc a den."

Myslili všichni, že jest nemocen a že blouzní z horečky.

I já jsem byl mezi těmi, kdož tak myslili.

II.

Obýval jsem malý útulný pokojík s oknem do ovocné a z části i květinové zahrady, v té době plné květů a vůně. Zde u otevřeného okna trávil jsem první ranní hodiny až do okamžiku, kdy Ludva, zbavena konečně nejnaléhavějších, jí přidělených povinností mladé hospodyně, nacházela chvíli oddechu, již věnovala pěstění svých květinových záhonů. Tu přicházela s konévkou v ruce a pokaždé zastavila se u mého okna, aby vyměnila se mnou pozdrav a se pozeptala jak pokračuje mé uzdravování. Naše ranní rozhovory bývaly velmi krátké. Přes to však, nedbal-li jsem o toto setkání a odešel o hodinu dříve na obvyklou svou dopoleni vycházku, tu vždy jakési neuspokojení naplnilo duši mou tesknym neklidem, jehož jsem se nezbavil, dokud jsem se opět nevrátil do domu svých hostitelů.

15

Fig. 2: A simple book page. TextBite has correctly identified that there are (segments of) two chapters, and linked the chapter header to the corresponding text. Note that page number is not a part of any logical segment.

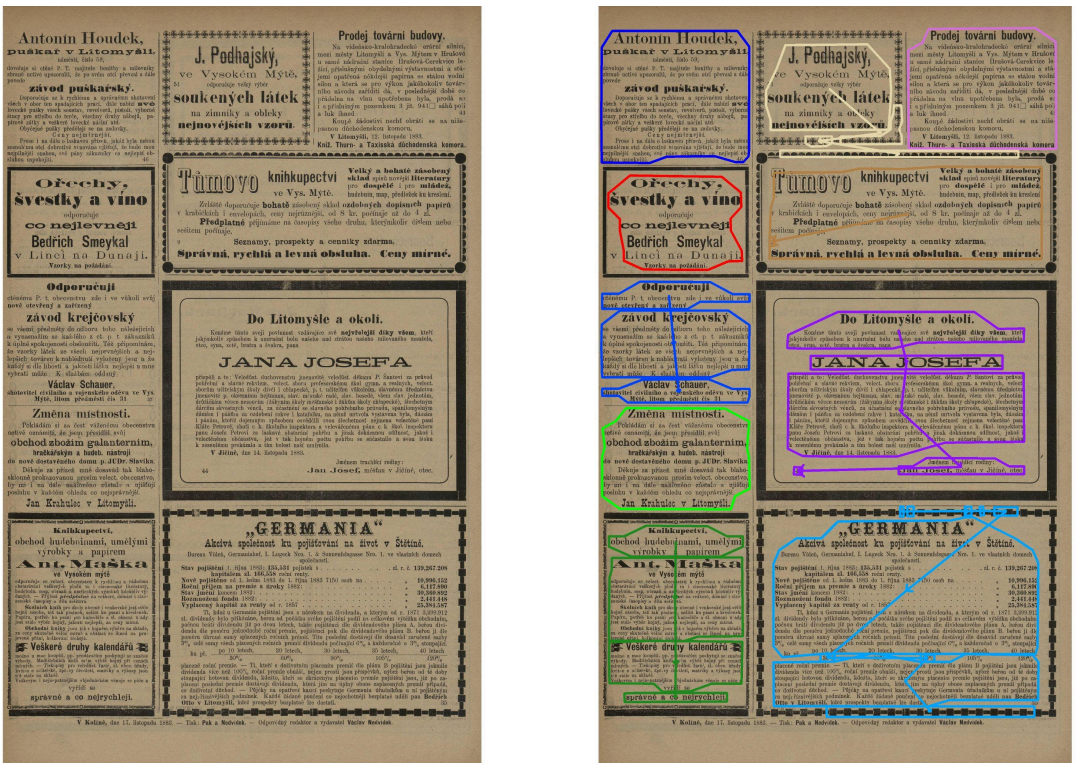


Fig. 3: A newspaper page with line advertisement. TextBite has correctly identified the individual ads. Note that it tries to stay as faithful as possible to the regions detected in the PAGE XML, including those that do not eventually correspond to actual text.

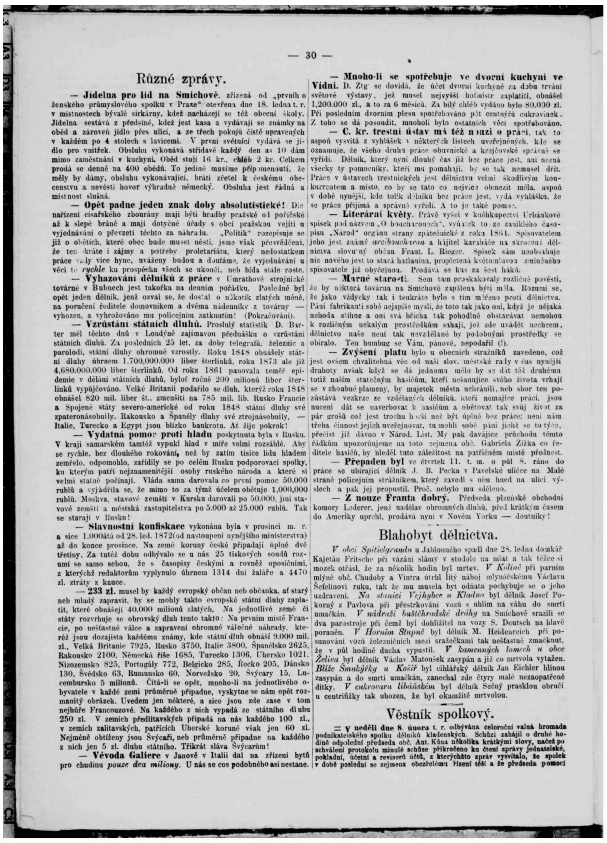
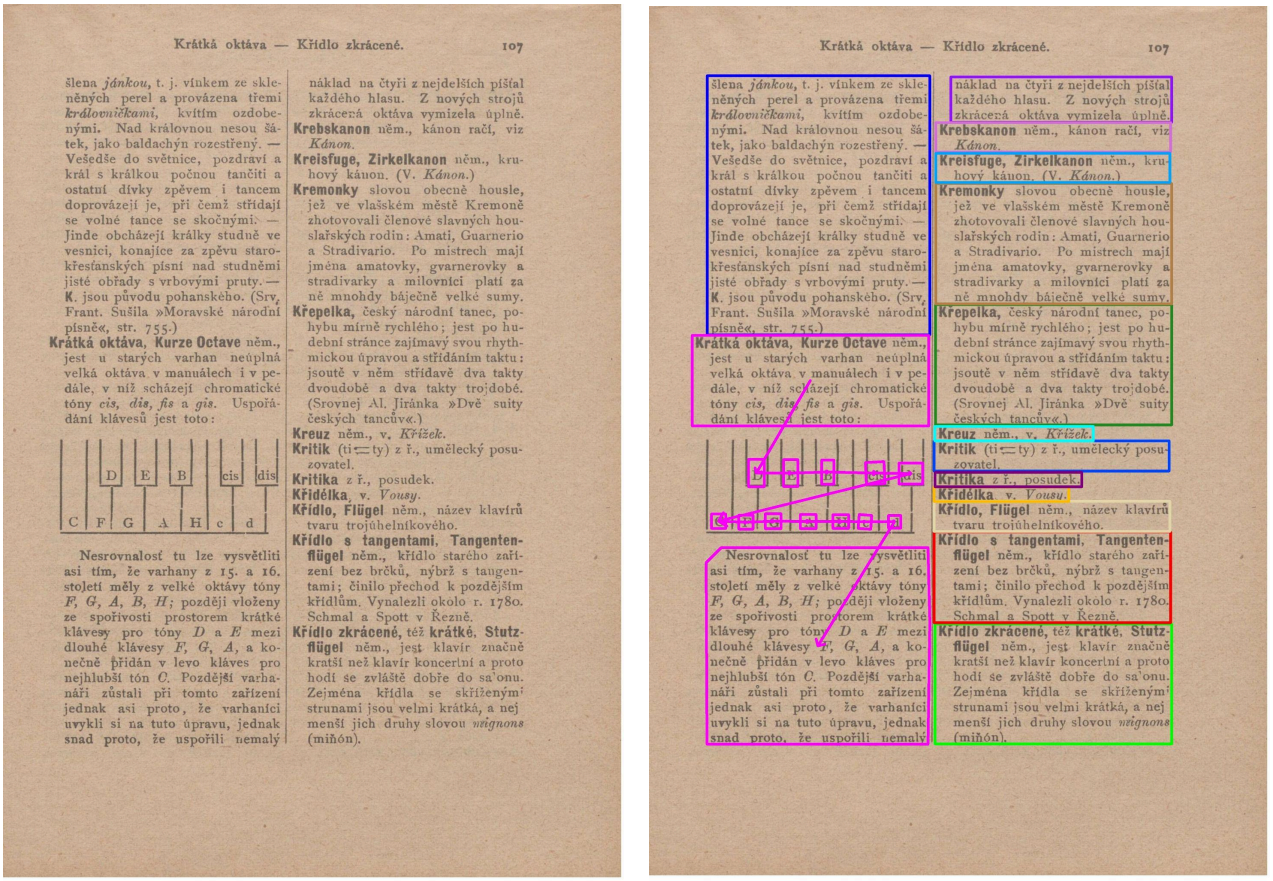


Fig. 4: Newspaper page of short news. Here, the original regions in the PAGE XML were collating several articles together, therefore TextBite had to split them. Note that the resulting regions (left column; upper part of right column) are simple rectangles, because the detection model in TextBite has not been trained to perform fine detection of text regions, allowing much faster operation.



šlena *jánkou*, t. j. vínkem ze skleněných perel a provázena třemi *královničkami*, kvítím ozdobenými. Nad královnou nesou šátek, jako baldachýn rozestřený. — Vešedše do světnice, pozdraví a král s králkou počnou tančiti a ostatní dívky zpěvem i tancem doprovázejí je, při čemž střídají se volné tance se skočnými. — Jinde obcházejí králky studně ve vesnici, konajíce za zpěvu starokřesťanských písní nad studněmi jisté obřady s vrbovými pruty. — K. jsou původu pohanského. (Srv. Frant. Sušila »Moravské národní písně«, str. 755.)

Krátká oktáva, Kurze Octave něm., jest u starých varhan neúplná velká oktáva v manuálech i v pedále, v níž scházejí chromatické tóny *cis*, *dis*, *fis* a *gis*. Uspořádání klávesů jest toto:

Nesrovnalosť tu lze vysvětliti asi tím, že varhany z 15. a 16. století měly z velké oktávy tóny *F*, *G*, *A*, *B*, *H*; později vloženy ze spořivosti prostorem krátké klávesy pro tóny *D* a *E* mezi dlouhé klávesy *F*, *G*, *A*, a konečně přidán v levo kláves pro nejhlubší tón *C*. Pozdější varhanáři zůstali při tomto zařízení jednak asi proto, že varhaníci uvykli si na tuto úpravu, jednak snad proto, že uspořili nemalý náklad na čtyři z nejdelších píšťal každého hlasu. Z nových strojů zkrácená oktáva vymizela úplně.

Krebskanon něm., kánon rací, viz *Kánon*.

Kreisfuge, Zirkelkanon něm., kruhový kánon. (V. *Kánon*.)

Kremonky slovou obecně housle, jež ve vlašském městě Kremoně zhotovovali členové slavných houslařských rodin: Amati, Guarnerio a Stradivario. Po mistrech mají jména amatovky, gvarnerovky a stradivarky a milovníci platí za ně mnohdy báječně velké sumy.

Křepelka, český národní tanec, pohybu mírně rychlého; jest po hudební stránce zajímavý svou rhythmickou úpravou a střídáním taktu: jsouť v něm střídavě dva takty dvoudobé a dva takty trojdobé. (Srovnej Al. Jiránka »Dvě suity českých tanců«.)

Kreuz něm., v. *Křížek*.

Kritik (ti=ty) z ř., umělecký posuzovatel.

Kritika z ř., posudek.

Křidélka, v. *Vousy*.

Křídlo, Flügel něm., název klavírů tvaru trojúhelníkového.

Křídlo s tangentami, Tangentenflügel něm., křídlo starého zařízení bez brků, nýbrž s tangentami; činilo přechod k pozdějším křídlům. Vynalezli okolo r. 1780. Schmal a Spott v Řezně.

Křídlo zkrácené, též **krátké, Stutzflügel** něm., jest klavír značně kratší než klavír koncertní a proto hodí se zvláště dobře do sa'onu. Zejména křídla se skříženými strunami jsou velmi krátká, a nejmenší jich druhy slovou *mignons* (miňón).

Fig. 5: An example page from a dictionary. TextBite has correctly identified the individual entries. Note that again, the irregular regions corresponding to individual notes were properly incorporated to the respective entry. Unfortunately on this page, TextBite has failed to link the end of the “Krátká oktáva” entry, which can be found in the right column.