sciendo
DOI 10.2478/jazcas-2023-0042

CAPEKDRACOR: A NEW CONTRIBUTION TO THE EUROPEAN PROGRAMMABLE DRAMA CORPORA

PETR POŘÍZKA
Department of Czech Studies, Faculty of Arts, Palacký University,
Olomouc, Czech Republic

POŘÍZKA, Petr: CapekDraCor: A New Contribution to the European Programmable Drama Corpora. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 244 – 253.

Abstract: The aim of this paper is to present the new *CapekDraCor* corpus and the *DraCor* project with its research-oriented concept of a programmable corpora focused on quantitative analyses within the framework of computational literary studies. This digital platform extends the possibilities of large-scale drama analysis with a focus on the dramatic character(s). The basic operationalisation is the interaction within a dramatic configuration, i.e., the scenic co-presence of two speakers, from which network data are automatically extracted, both global networks of interactions of dramas and data characterising individual actors, i.e., literary characters.

The paper demonstrates the *CapekDraCor* corpus, a new contribution to the extensive *DraCor* database, and presents the way the data are processed with respect to their specific multi-layered structure. The corpus contains all the plays written by Karel and Josef Čapek and the data are processed in a standardized format based on XML and general TEI guidelines for processing drama with a defined basic drama tagset. *CapekDraCor* also uses the newly created EZdrama format for data processing, which works as an intermediate step from .txt to .xml file as a lightweight YAML-like markup language. A file in this format can be automatically converted into a DraCor-ready XML file with a TEI header.

The advantage of the programmable corpora concept is the possibility to use suitably structured data for drama research outside the *DraCor* platform and with other methods or tools for textual analysis. Simultaneously, this approach moves the researcher from the technical requirements of the analysis to operationalised computational analysis based on research questions and pre-prepared and flexible tools. *DraCor* is a unique open infrastructure (both in terms of data and tools) for the analysis of European drama, currently comprising 15 corpora in 10 different languages with a total of about 3,000 plays from a wide range of periods.

Keywords: data annotation, computational literary studies, corpus building, drama, DraCor, network analysis, quantitative analysis

1 INTRODUCTION AND STATE OF THE ART

The Čapek brothers and their plays represent a unique literary, linguistic and socio-cultural phenomenon, because of the interplay between literary fiction and strong moral, political and social commentary in their plays. Karel Čapek is a world-famous writer and his theatrical work undoubtedly represents one of the

most important parts of his entire oeuvre. His brother Josef, who co-wrote three plays and was himself the author of two other dramas, also contributed significantly to this work. In total, the dramatic work of the Čapek brothers comprises ten plays and has been discussed many times in the scholarly literature and analysed from various perspectives: theatrical, literary and linguistic. Philological analyses have thus far relied mainly on the introspection of the researcher and rather traditional, structuralist interpretative approaches without the use of computer text analysis. To this day, the plays by Josef have remained largely unexplored by researchers, including linguists and literary scholars with respect to quantitative analyses. The *CapekDraCor* corpus, which we present here, aims to change this situation and enable researchers to analyse these texts using modern methods of corpus and quantitative linguistics within the framework of Computational Literary Studies. Although the *Czech National Corpus* (CNC) already has a corpus of Karel Čapek's plays (Čermák et al. 2007) called *capek*, its processing is limited because it does not reflect their multilayered structure (the textual and metatextual parts are not separated), which can (and does) affect subsequent quantitative analyses and their results. See Fig. 1 below, where both the comments and the characters' labels preceding the dialogues, or even the initial cast list, are represented as text in the CNC *capek* corpus (Fig. 1):



Fig. 1. The Czech National Corpus and Karel Čapek's dramas: text and meta-text are mixed and not differentiated in the CNC *capek* corpus.

Thus, the use of modern and exact methods of computer text analysis requires the re-processing of Čapek's plays into a new annotated corpus that will reflect the specifics of the plays, including their multi-layered structure, which will enable the analysis not only of the text as a whole, but also of its sub-parts (by separating the dialogue layer from the metatextual parts). In addition, other structural elements of the text should be possible to use as a unit or parameter of analysis, and the literary

characters of the individual plays are especially the key in this respect. These possibilities are offered by the *DraCor* project focusing on European drama, which seeks to apply computational methods to digitized literary texts, in line with a similar COST Action on European novels (Schöch et al. 2018).

2 DRACOR AS A DIGITAL PLATFORM FOR THE DRAMAS

DraCor is an open platform as well as a growing network of textual resources for hosting, accessing, and analysing European dramas, which also includes tools designed for data mining. It currently includes 15 corpora in 10 different languages with a total of approximately 3,000 plays of a wide temporal range.

The *DraCor* infrastructure is based on the concept of "Programmable Corpora", a new term proposed in Fischer et al. (2019) for research-oriented corpora providing APIs, which is an important methodological concept that moves the researcher from the technical requirements of analysis to an operationalised computational analysis that is based on research questions and pre-prepared and flexible tools that enable such analysis. I quote: "Projects like DraCor seek to provide the digital literary studies with a reliable and extensible infrastructure so that the research community can focus on research questions. (…) Programmable Corpora facilitate the implementation of research questions around corpora. (…) and make it easier to decide on which level of the platform your own research process starts." (cf. Fischer et al. 2019). Another important conceptual term related to the DraCor base is vanilla corpora: "Similar to Paul Fièvre's collection 'Théâtre classique', these corpora are designed as vanilla corpora, which initially contain hardly any special markup beyond the necessary, but are freely available and can therefore be forked, enriched and expanded" (ibidem).[1] The necessary annotation therefore concerns specific aspects of dramatic texts.

2.1 Technical specification

DraCor data are processed in a standardised format based on XML and general TEI requirements or guidelines (P5 model) for drama processing with a defined bas ic drama tagset (TEI Performance Texts 2023), bu t texts are also available in other (different) formats to suit the needs of individual tools and analyses (CSV, JSON, GEXF, GraphML, RDF).

Additional corpora can be easily linked to the *DraCor* infrastructure, and all existing data mining and visualisation methods of the platform are easily applicable to newly added corpora. The only prerequisite is that they are encoded in TEI/XML (see

[1] This does not mean, however, that it is impossible to additionally annotate the texts linguistically if necessary: the Shakespearian corpus *ShakeDraCor* is even linguistically annotated (lemma and tag by Morphadorner and NUPOS English Tagset).

above) and freely available as a corpus hosted on GitHub that can be cloned and loaded directly into a structured database. *DraCor* relies on eXist-db as XML database to process TEI files and provide functions for researching the corpora. The frontend is built using *React* (https://reactjs.org/), which is responsive and easily extensible, with the focus on APIs. Other tools in the system are *SPARQL* (an application for eXist-db), *ezlinavis* (a didactic tool), *ShinyDraCor* (an R-based framework that allows interactive visuali s ations to be displayed in the browser, relying entirely on the *DraCor* API for data retrieval), and most recently *rdracor* (a new R interface for the *DraCor* API) (see https://dracor.org/ in the Tools section for more information).

3 CAPEKDRACOR CORPUS AND DATA PROCESSING

The *CapekDraCor* corpus as a new contribution to the *DraCor* corpora is composed of the plays written by Karel and Josef Čapek, namely (1) *Loupežník* (‘The Robber ’, 1920); (2) *R.U.R.* (1920); (3) *Věc Makropulos* (‘The Makropulos Affair’, 1922); (4) *Bílá nemoc* (‘The White Disease’, 1937); (5) *Matka* (‘The Mother’, 1938); (6) *Lásky hra osudná* (‘The Fateful Game of Love’, 1910); (7) *Ze života hmyzu* (‘The Insect Play’, 1921); (8) *Adam Stvořitel* (‘Adam the Creator’, 1927); (9) *Země mnoha jmen* (‘The Land of Many Names’, 1923); (10) *Gassirova loutna* (‘Gassir’s Lute’, 1923). As far as the authorship of the individual plays is concerned, Karel wrote plays Nos. 1–5, together with his brother Josef they wrote plays Nos. 6–8, and Josef is the author of plays Nos. 9–10. Critical editions of the plays of the Čapek brothers from the following editions were used as a textual basis:[2]

- Karel Čapek: *Spisy Karla Čapka: Dramata*, Československý spisovatel 1992/94 (editor Emanuel Macek); dramas: *Loupežník*, *R.U.R.*, *Věc Makropulos*, *Bílá nemoc*, *Matka;* with respect to the edition Česká knižnice, Brno: Host 2014 (editor Jiří Holý); dramas: *Loupežník*, *R.U.R.*, *Bílá nemoc*
- Karel Čapek – Josef Čapek: *Ze společné tvorby*, Praha: Československý spisovatel 1982 (editor Emanuel Macek); dramas: *Lásky hra osudná*, *Ze života hmyzu*, *Adam Stvořitel*
- Josef Čapek: *Beletrie 1*, Brno: Triáda 2011 (editor Daniel Vojtěch); dramas: *Země mnoha jmen*, *Gassirova loutna*

These plays were processed into standardised, commonly used and extended formats (.txt, .xml) and are thus not limited to use in the *DraCor* database, which will increase their subsequent usability by other researchers and other software tools (e.g. R and packages like *stylo*, *quadrama* etc.). The following categories of metadata were encoded (and can subsequently be used for analyses):

[2] All texts are freely available and copyright is not applicable (copyright protection expires 70 years from the author’s death). Other texts for the purposes of this project (with subsequent publication of the results: corpora and papers) were kindly provided free of charge by the publishing houses Host (with the consent of the editor J. Holý) and Triáda (also with the consent of the editor D. Vojtěch).

- external and textual: *author*, *title*, *year (written, printed, premiered)*, *language*, *genre*, *edition; cast list*;
- internal: text structure (*prologue/act/scene/epilogue*); layer distinction (*characters/dialogues/commentary/stage directions*); literary characters (*cast group* and *role* ; *gender/sex*).

The texts have not yet been annotated linguistically (lemma, part-of-speech category, morphological tags), but this variant is being considered for processing for the Czech National Corpus.

The final and most important format is undoubtedly the TEI/XML-encoded version of texts. The processing is carried out, however, through an intermediate stage – conversion to *EZdrama* format, which makes it easier and faster to create the required XML file. See the diagram below (Fig. 2):



Fig. 2. Individual stages of text processing for the *DraCor* database

EzDrama format is a lightweight Markdown-like (or YAML-like) markup language and a file in this format can be created either manually or semi-automatically through search–replace conversions using regular expressions. This phase is followed by automatic conversion to DraCor ready XML file with TEI header.

There will now be a demonstration of what the text in *EZdrama* format looks like before conversion to XML (syntax and example below). Several metasymbols and line breaks are used to encode the text, namely: [3]

means a structural part like an *act* (#) or *scene* (##); technically the number of nesting levels is not limited

$ means a new *stage direction* or *commentary* ; brackets () are also used and automatically converted to stage directions or comments – they do not require any special editing

@ means the text line contains the name (label) of the character preceding the dialogue; note: any other line without these special symbols will be treated as the direct speech of the last encoded speaker (@)

~ means that this and the following unmarked lines are in verse (poet ry)

Let us look at a short fragment of dialogue between Mimi and Fanka from the play *Loupežník* (The Robber), at the beginning of the second act (DĚJSTVÍ DRUHÉ), prelude (PŘEDEHRA) – Fig. 3:

[3] An overview of all metasymbols and more information about the EZdrama format can be found on the website https://github.com/dracor-org/ezdrama/blob/main/README.md.

```
#DĚJSTVÍ DRUHÉ
##PŘEDEHRA
$Loupežník se zavázanou hlavou a Doktor vycházejí z hospody U Beránka, sednou
si venku u stolu a mastí karty.
@MIMI (vyjde z vrat s nějakým šitím):
Proč nikoho neposílá? Co se s ním děje? Bože, slituj se nade mnou!
@FANKA (vyjde z vrat):
Nikam nechodila, Mimi!
@MIMI:
Já nikam nejdu. (Sedne si na lavičku.) Ach, kdyby už poslal zprávu, jak mu je!
```

Fig. 3. EZdrama format: lines are marked with special symbols at the beginning; an excerpt from the play *Loupežník* (The Robber)

Specific parts marked with special symbols are colour-coded for better readability (comments or stage directions in green, character names in blue, acts/scenes in red). The texts are automatically converted from this EZdrama format to the DraCor-ready XML file – cf. Fig. 4:

```
<div type="act">
  <head>DĚJSTVÍ DRUHÉ</head>
  <div type="scene">
    <head>PŘEDEHRA</head>
    <stage>Loupežník se zavázanou hlavou a Doktor vycházejí z hospody U
    Beránka, sednou si venku u stolu a mastí karty.</stage>
    <sp who="#mimi">
      <speaker>MIMI:</speaker>
      <stage>(vyjde z vrat s nějakým šitím)</stage>
      <p>Proč nikoho neposílá? Co se s ním děje? Bože, slituj se nade
      mnou!</p>
    </sp>
    <sp who="#fanka">
      <speaker>FANKA:</speaker>
      <stage>(vyjde z vrat)</stage>
      <p>Nikam nechodila, Mimi!</p>
    </sp>
    <sp who="#mimi">
      <speaker>MIMI:</speaker>
      <p>Já nikam nejdu. <stage>(Sedne si na lavičku.)</stage> Ach,
      kdyby už poslal zprávu, jak mu je!</p>
    </sp>
```

Fig. 4. DraCor-ready XML file after the automatic conversion from EZdrama format; an excerpt from the play *Loupežník* (The Robber)

The TEI header contains the necessary metatextual information, an overview of literary characters and their characteristics (cast list and role description), plus other necessary meta-information (see above).[4] The predominant way of processing plays in

[4] For more information about the TEI header see the website: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html, or https://teibyexample.org/tutorials/TBED02v00.htm.

the *DraCor* subcorpora (TEI/XML) (as shown in Fig. 4) is based on adequate segmentation of the dramas mainly through XML elements into different layers of text/meta-text without any additional linguistic annotation, or through XML attributes used to encode other aspects. The distinction is mainly related to spoken dialogues (<p> element for prose and <l> element for poetry) versus comments and stage directions (<stage> element); <sp> i s the element to mark up an individual speech in a dramatic text (the actual utterances) and <speaker> elements are used to mark up dramatic characters (producing the speech); element and *type* attribute tags with values *act*, *scene* (and others) are used to indicate the text structure. [5]

4 DIGITAL NETWORK ANALYSIS OF THE DRAMAS

The *DraCor* platform approach aims to extend the possibilities for large-scale drama analysis with a focus on the dramatic character(s). The literary figure is understood here as the central element of narrative (Todorov 1977; Fořt 2008). A network-analytic conceptualisation of dramatic interaction is used as the theoretical basis (Moretti 2011): the basic operationalisation is the *interaction* within a dramatic configuration, which is the scenic co-presence of two speakers (Trilcke 2015). Based on this notion of relationships, network data are automatically extracted, both global networks of ‚interactions' of the dramas (density, average degree, connectivity, etc.) and data characterising individual actors, i.e., literary figures (degree and various other centrality indices). The following picture from the *CapekDraCor* corpus illustrates the network of relationships between the characters in Karel Čapek's drama R.U.R. (Fig. 5):

As mentioned above, the role of the character in relation to other participants is of key importance regarding the interpretation of narrative. It is a co-presence network that is based only on the co-occurrence (or co-appearance) of the characters in each section (in this case, scene). The size of the bubble is the weighted degree of the node (i.e., the figure), which is the sum of the edge weights for the edges incident to that node.

*DraCor*system allows such networks to be generated based on various criteria – all characters, selected characters (e.g. male or female, only 'main'figures, etc.) and offers various metrics that can be used for further analysis. A major asset is the already quite robust database of plays by various European playwrights with a large time span, which can again be used for comparisons and contrastive analyses (via *DraCor*metrics). For illustrative examples of such comparative analyses see e.g. Skorinkin et al. (2018).

[5] For a more detailed description of the XML format related to dramas cf. https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html, https://teibyexample.org/tutorials/TBED05v00.htm2, or Base Tagset for Drama: https://tei-c.org/Vault/P4/doc/html/DR.html.

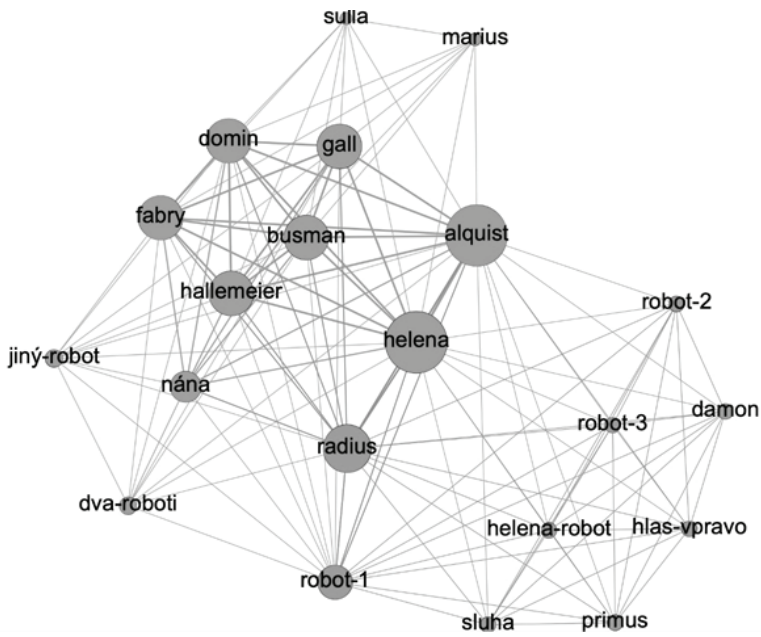


Fig. 5. *CapekDraCor*: co-occurrence network of characters from Karel Čapek's R.U.R.

5 CONCLUSIONS AND POSSIBILITIES OF USING PROGRAMMABLE CORPORA

Case studies (or examples) can be found in the collection of research papers on drama analysis using data provided through the *DraCor*database and using the *Dracor*tools on the project website.[6]

The *DraCor*infrastructure provides a unique, data-rich and extensive base not only for the analysis of the individual plays or authors, but also for comparative or contrastive analyses, for which other European playwrights and their works can be used (see the relevant *DraCor*subcorpora). It is thus possible to trace the development of selected aspects of European drama over time (plays from the 16th to 20th centuries) as well as other narratological aspects of different European playwrights, the complexity of dramas (i.e., the networks' complexity of relationships between characters) or sub-aspects such as the development of commentary sections, the ratio and characteristic of male and female figures, the role and contribution of characters to the narrative of the text, etc. An indisputable advantage with great research potential is the overall concept of programmable corpora and the possibility to use suitably structured data (open access source) for drama research outside the

[6] See the website: https://dracor.org/doc/research.

*DraCor*platform and with other methods or tools for textual analysis. The most illustrative example is the *QuaDramA*case study, which uses the *GerDraCor*corpus for illustrative drama analysisand demonstrates the flexibility and versatility of the programmable corpora approach (Reiter – Pagel 2022).

In quantitative research, analyses include both descriptive statistics and an exploratory quantitative approach (inferential statistics), in this case, for example, based on network analysis of dramas. The *CapekDraCor*corpus, as well as the entire *DraCor*infrastructure that we wanted to present in this text, are designed to meet these needs and objectives.

ACKNOWLEDGEMENTS

The research has been supported by the Program of the European Commission funded project CLS INFRA (https://clsinfra.io/) from the European Union's Horizon 2020 program (grant agreement No. 101004984).

References

Čapek, J. (2011). Beletrie 1. Praha: Triáda.

Čapek, K., and Čapek, J. (1982). Ze společné tvorby. Praha: Československý spisovatel.

Čapek, K. (1992/94). Spisy Karla Čapka: Dramata. Praha: Československý spisovatel.

Čapek, K. (2014). Tři hry (Loupežník, RUR, Bílá nemoc). Brno: Host.

Čermák, F. et al. (2007). Capek: korpus vlastních textů Karla Čapka. Praha: Ústav Českého národního korpusu. Accessible at: http://www.korpus.cz.

DraCor. Drama Corpora Project [database]. Accessible at: https://dracor.org/.

Reiter, N., and Pagel, J. (2022). Analysis of Dramatic Texts. Package DramaAnalysis. Accessible at: https://quadrama.github.io/DramaAnalysis/tutorial/3/index.html.

Fischer, F. et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht University.

Fořt, B. (2008). Literární postava. Vývoj a aspekty naratologických zkoumání. Praha.

Herman, D. (2002). Story Logic. Problems and Possibilities of Narrative. University of Nebraska Press.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. Mots, 1, pages 127–165.

Moretti, F. (2011). Network Theory, Plot Analysis. Stanford Literary Lab Pamphlets 2. Accessible at: http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf.

QuaDramA. Quantitative Drama Analytics [Software]. Accessible at: https://github.com/quadrama.

React. The Library for Web and Native User Interfaces. Accessible at: https://reactjs.org/.

The Shakespearian corpus ShakeDraCor. Accessible at: https://dracor.org/shake.

Schöch, Ch. et al. (2018). Distant Reading for European Literary History. A COST Action [Poster]. In DH2018: Book of Abstracts / Libro de resúmenes. Mexico: Red de Humanidades Digitales A. C. Accessible at: https://dh2018.adho.org/en/?p=11345.

Skorinkin, D., Fischer, F., and Palchikov, G. (2018). Building a Corpus for the Quantitative Research of Russian Drama: Composition, Structure, Case Studies. In

Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference »Dialogue 2018«, pages 662–682, Moscow.

stylo: Stylometric Multivariate Analyses (version 0.7.4). [Software]. Accessible at: https://cran.r-project.org/web/packages/stylo/index.html.

TEI Performance Texts (2023). In TEI P5 – Guidelines for Electronic Text Encoding and Interchange (version 4.6.0). Accessible at: https://tei-c.org/release/doc/tei-p5-doc/en/html/DR.html.

Textometrie project: TXM (version 0.8.1) [Software]. Accessible at: http://textometrie.enslyon.fr.

The R Project for Statistical Computing: R (version 4.0.4) [Software]. Accessible at: https://www.r-project.org/.

Todorov, T. (1977). The Poetics of Prose. Ithaca – New York: Cornell university Press.

Trilcke, P., Fischer, F., and Kampkaspar, D. (2015). Digital Network Analysis of Dramatic Texts. In DH2015: »Global Digital Humanities«. University of Western Sydney.