Resource reviewed	
Title	InterCorp
Editors Bibliographical description of responsible personnel. Please indicate names as “forename surname”.	Alexandr Rosen, Martin Vavřín, Adrian Zasina
URI	http://www.korpus.cz/
Publication Date Format should be either yyyy, yyyyff. or yyyy-yyyy, e.g. 2007 or 2007-2013	2008-2017
Date of last access	09.03.2018

Reviewer	
Surname	Kim
First Name	Agnes
Organization	Institut für Slawistik, Universität Wien
Place	Vienna
Email	agnes.kim (at) univie.ac.at

Personnel

Editors	Alexandr Rosen Martin Vavřín Adrian Zasina

General Information		
Bibliographic description	Can the text collection be identified in terms similar to traditional bibliographic descriptions (title, responsible editors, institution, date(s) of publication, identifier/address)? (cf. Catalogue 1.1)	✓
Contributors	Are the contributors (editors, institutions, associates) of the project documented? (cf. Catalogue 1.3)	✓
Contacts	Is contact information given? (cf. Catalogue 1.4)	✓
Aims		
Documentation	Is there a description of the aims and contents of the text collection? (cf. Catalogue 2.1)	✓
Purpose **We assume that the text collection is published to be used in a certain context and that the intended usage scenario can be inferred (if it is not described), so if the purpose of the text collection is not stated explicitly, please answer according to your own impression.**	What is the purpose of the text collection? (cf. Catalogue 2.2)	Research, Teaching, General purpose, other: translation
Kind of research **In many cases both qualitative and quantitative research will be possible. Please choose the methodological orientation that is prevailing form your point of view.**	What kind of research does the collection allow to conduct primarily? (cf. Catalogue 3.1.8)	Qualitative research

Self-classification	How does the text collection classify itself (e.g. in its title or documentation)? (cf. Catalogue 2.3)	Corpus
Field of research	To which field(s) of research does the text collection contribute? (cf. Catalogue 2.2)	Linguistics, other: Translation studies
Content		
Era **Classics: before 500 CE. Medieval: 501 CE until 1500 CE. Early Modern: 1501 CE until 1800 CE. Modern: 1801 CE until 1945. Contemporary: 1945 until today.**	What era(s) do the texts belong to? (cf. Catalogue 2.5)	Contemporary
Language **Please choose the language name(s) that correspond(s) best to the language(s) of the texts, e.g. for Old English choose English; for Mexican Spanish choose Spanish. Choose 'other' if none of the given language names matches the language(s) in question. If you wish to specify the language(s) further, you can give an additional explanation in the 'note' field.**	What languages are the texts in? (cf. Catalogue 2.5)	Arabic, Chinese, Danish, English, Finnish, French, German, Greek, Hebrew, Hindi, Italian, Japanese, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish, Turkish, other: Albanian, Belarusian, Bulgarian, Catalan, Croatian, Czech, Dutch, Estonian, Hungarian, Icelandic, Latvian, Lithuanian, Macedonian, Malay, Maltese, Romani, Romanian, Serbian, Slovak, Slovene, Ukrainian, Vietnamese

Types of text **Literary works: artistic creations like poems, novels, plays, transmitted in one or several documents.** **Private documents: personal or family documents or papers, e.g. letters and diaries.** **Essays: formal or informal, argumentative or reflective writings.** **Newspaper/journal articles: articles discussing recent news of general interest or a specific topic, published in a newspaper or journal.** **Charters: texts documenting a legal fact by using a special form supporting its validity, e.g. the Magna Carta.** **Inscriptions: texts on stone, metal, wood, etc., fixed through carving, engraving or embossing.** **Files/records: classified collections of documents, e.g. business and trial records, personnel and administration files.** **Protocols: written records of meetings.** **Scientific papers: (previously published) scientific literature, e.g. articles, monographs or reference works.** **Speech transcripts: written records of speech, transcribed orthographically or phonetically, for example.**	What kind of texts are in the collection? (cf. Catalogue 2.5)	Literary works, Newspaper/journal articles, Charters, Protocols, other: Subtitles, Bible translations
Additional information **Information that is derived from the texts (e.g. analytical data or visualizations) is not considered here.** **Commentary: (scholarly) commentary on the content of the documents or regarding other textual phenomena.** **Context material: additional (textual) sources which put the texts in the collection into context but are not considered part of the collection itself.** **Facsimiles: any copy of historical documents.**	What kind of information is published in addition to the texts? (cf. Catalogue 2.5)	none
Composition		
Documentation	Are the principles and decisions regarding the design of the text collection, its composition and the selection of texts documented? (cf. Catalogue 3.1.1-3.1.3)	✓

Selection **Selection criteria: factors that guided the selection of texts for the collection and the composition of the text collection.** **Language: one or several standard languages or other language varieties (including for example sociolects).** **Author: one or several persons or figures that are supposed to have authored the texts.** **Country: one or several countries or geographical regions.** **Epoch: one or several (historical) epochs or other kinds of time periods.** **Genre: literary (e.g. novella, fabel) or non-literary genres/text types (letter, statute, recipe, chat log, etc.).** **Topic: one or several thematic aspects of the texts.** **Style: one or several writing styles characteristic for authors, periods, schools, text types, etc. (e.g. romantic or realist novels, satirical texts, Victorian style).** **Linguistic characteristics: language properties other than style (e.g. the presence of certain grammatical constructions).**	What selection criteria have been chosen for the text collection? (cf. Catalogue 3.1)	Language, Epoch, other: available Czech bitext
Size If the size of the text collection is not given explicitly but can be inferred, please choose appropriate numbers, otherwise choose 'unknown'.		
Texts/records	How large is the text collection in number of texts/records? (cf. Catalogue 3.1.4)	> 1000
Tokens **Tokens: Sequences of strings delimited by whitespace or punctuation and roughly corresponding to words.**	How large is the text collection in number of tokens? (cf. Catalogue 3.1.4)	> 10 Mio.
Structure	Does the text collection have identifiable sub-collections or components? (cf. Catalogue 3.1.5)	✓
Data acquisition and integration		
Text recording	Does the text collection record or transcribe the textual data for the first time? (cf. Catalogue 3.1.6)	✗

Text integration	What kind of material has been taken over from other sources? (cf. Catalogue 3.1.6)	Full texts, Metadata, Annotations
Quality assurance **Choose 'yes' if there has been a quality check for which results are reported, regardless of whether corrections have been made or not.**	Has the quality of the data (transcriptions, metadata, annotations, etc.) been checked? (cf. Catalogue 3.1.7)	✓

Typology **General purpose collection: a text collection of a very general nature (e.g. Wikisource, Project Gutenberg); often created in a collaborative fashion; with no specific or very loose selection criteria; usually not bound to a certain time frame for its creation and completion.** **Corpus: a collection of texts that has been created according to some selection criteria (language, author, country, epoch, genre, topic, style, etc.) which makes it more specific than a general purpose collection; not necessarily aiming at completeness or representativeness; e.g. the 'Corpus of English Religious Prose', 'Letters of 1916', 'Corpus of Literary Modernism'.** **Collection of records: a collection of texts that are held together out of organisational reasons, e.g. a collection of historical documents that has been kept in the same archive.** **Canon: collection of works that is considered most important for a certain period, culture or discipline (e.g. the biblical canon, the canon of English 19th century literature); might be formally approved or authoritative and subject to debate and revision.** **Complete works/œuvre: collection of all works by a single author (e.g. complete works of Mark Twain).** **Reference corpus: collection of texts that have been selected in order to be representative for a certain genre or language (e.g. reference corpus of New High German Language).** **Contrastive corpus: a collection of texts aiming at the systematic comparison of its sub-components, to get to a description of differences and similarities between them (e.g. FinDe, a contrastive corpus of Finnish and German).** **Parallel corpus: a collection of texts which are contrasted with other versions, often translations (e.g. the Parallel Bible Corpus). A parallel corpus can be considered a certain kind of contrastive corpus.** **Diachronic corpus: a collection of texts that have been selected in order to reflect evolution over time (e.g. the Diachronic Corpus of Present-Day Spoken English (c. 1960-1980)).**	Considering aims and methods of the text collection, how would you classify it further? For definitions please consider the help-texts. (cf. Catalogue 3.1.8)	Reference corpus, Parallel corpus
Data Modelling		

Text treatment **Normalized transcription: if the orthography has been normalized according to a chosen standard (e.g. 'seyn' to 'sein').** **Orthographic transcription: a transcription that employs the standard spelling system of each target language (e.g. the surname "Pushkin" in English orthographic transcriptions of the Russian surname "Пу́шкин").** **Phonetic/phonemic transcription: a transcription that is the visual representation of speech sounds or phones (e.g. [ˈpuʂkʲɪn]) or a phonemic transcription (e.g. /ˈpʊʃkɪn/).** **Diplomatic transcription: a transcription of the document taking into account features like spelling, punctuation, abbreviations, deletions, insertions, alterations, etc.** **Transliteration: A conversion of a text from one script to another (e.g. "Russia" in Cyrillic script, "Россия", is transliterated as "Rossiya" in Latin script).** **Edited text: A reading text as constituted by the editor(s), based on text-critical procedures like recensio, examinatio, emendatio, correction, normalization, modernization etc.** **Translated text: Any translations into languages different from that of the original text.** **Summarized text: A summary of the source text.** **Sampled text transcriptions: parts of texts that have been selected and transcribed to represent whole texts (e.g. out of theoretical considerations or for statistical reasons).**	How are the textual sources represented in the digital collection? (cf. Catalogue 3.2.1)	other: according to the original sources
Basic format **Plain text: a pure sequence of character codes supported by the underlying standard (ASCII, Unicode).** **XML: Extensible Markup Language, a general markup language that defines a set of rules for encoding documents.** **HTML: Hypertext Markup Language, a standard markup language for web pages.**	In which basic format are the texts encoded? (cf. Catalogue 3.2.4)	XML
Annotations		

Annotation type **Semantic annotations: e.g. key words, links to (controlled) vocabularies, norm data.** **Linguistic annotations: additional information about linguistic characteristics of the texts, e.g. lemmata or PoS-tags.** **Editorial annotations: e.g. editorial comments and/or text-critical components such as the apparatus criticus.** **Structural information: e.g. markup to capture the textual structure (e.g. headings, chapters) and layout information (e.g. paragraphs, indents).**	With what information are the texts further enriched? (cf. Catalogue 3.2.2)	Linguistic annotations, Structural information
Annotation integration **Please choose 'not applicable' if there are no annotations.**	How are the annotations linked to the texts themselves? (cf. Catalogue 3.2.2)	Embedded
Metadata		
Metadata type **Descriptive: to describe and identify a resource, e.g. unique identifier, physical, bibliographic and content related attributes (such as medium, dimensions, author, title, publication year, genre, topic).** **Structural: information about the internal structure of a resource (such as parts, volumes, chapters, sections, pages).** **Administrative: for example technical details, access rights, history of changes.**	What kind of metadata are included in the text collection? (cf. Catalogue 3.2.3)	Descriptive, Structural, Administrative
Metadata level	On which level are the metadata included? (cf. Catalogue 3.2.2)	Collection parts/components, Individual texts
Data schemas and standards		
Schemas **General standardized schema: TEI All, TEI Lite, TCF, EAD, etc.** **Customized standard schema: a project specific customization of a standardized schema, e.g. a certain RDFS(chema) or the DTABf.** **Project specific schema: a schema that does not conform to any standard vocabulary, e.g. a custom XML dialect.**	What kind of data/metadata/annotation schemas are used for the text collection? (cf. Catalogue 3.2.4)	General standardized schema, Customized standard schema

Standards **TEI: Text Encoding Initiative, cf. http://www.tei-c.org** **CEI: Charters Encoding Initiative, cf. https://www.cei.lmu.de** **EAD: Encoded Archival Description, cf. https://www.loc.gov/ead/** **(X)CES: Corpus Encoding Standard (in XML), cf. https://www.cs.vassar.edu/CES/ and http://www.xces.org/** **Dublin Core: a set of vocabulary terms for the description of web resources; cf. http://dublincore.org/** **EDM: Europeana Data Model, cf. http://pro.europeana.eu/page/edm-documentation** **METS: Metadata Encoding and Transmission Standard, cf. http://www.loc.gov/standards/mets/** **MODS: Metadata Object Description Schema, cf. www.loc.gov/mods/** **SKOS: Simple Knowledge Organization System, cf. https://www.w3.org/2004/02/skos/** **OWL: Web Ontology Language, cf. https://www.w3.org/OWL/** **IMDI: Isle Metadata Initiative, cf. https://tla.mpi.nl/imdi-metadata/** **CMDI: Component Metadata Infrastructure, cf. https://www.clarin.eu/content/component-metadata** **TCF: Text Corpus Format, cf. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format** **OLAC: Open Language Archive Metadata, cf. http://www.language-archives.org/OLAC/metadata-20080531.html** **EAGLES: Guidelines of the Expert Advisory Group on Language Engineering Standards, cf. http://www.ilc.cnr.it/EAGLES/browse.html** **standardized PoS tagset(s): Part-of-Speech tagsets that have been standardized, for example the 'Part-of-Speech Tagging Guidelines for the Penn Treebank Project'.**	Which standards for text encoding, metadata and annotation are used in the text collection? (cf. Catalogue 3.2.4)	standardized PoS tagset(s)
Provision		
Accessability of the basic data	Is the textual data accessible in a source format (e.g. XML, TXT)? (cf. Catalogue 4.1)	✗

Download	Can the entire raw data of the project be downloaded (as a whole)? (cf. Catalogue 4.2)	✗
Technical interfaces **OAI-PMH: Protocol for Metadata Harvesting; a protocol for harvesting metadata descriptions of items in a collection.** **REST: Representational State Transfer; a paradigm for the architecture of so-called RESTful web services.** **SPARQL endpoint: a SPARQL Protocol and RDF Query Language endpoint to retrieve data stored in the RDF format.** **General API: an Application Programming Interface other than OAI-PMH or REST.**	Are there technical interfaces which allow the reuse of the data of the text collection in other contexts? (cf. Catalogue 4.2)	none
Analytical data	Besides the textual data, does the project provide analytical data (e.g. statistics) to download or harvest? (cf. Catalogue 4.3)	✗
Reuse	Can you use the data with other tools useful for this kind of content? (cf. Catalogue 4.4)	✓
User Interface		
Interface provision **For example, a website created for the presentation of the texts or a software developed for the display and usage of the text collection in question is considered a dedicated user interface, while a general repository (e.g. a library publication server), versioning platform (e.g. GitHub) or archive (e.g. Zenodo) is not.**	Does the text collection have a dedicated user interface designed for the collection at hand in which the texts of the collection are represented and/or in which the data is analyzable? (cf. Catalogue 5.1)	✓
User Interface questions		
Usability	From your point of view, is the interface of the text collection clearly arranged and easy to navigate so that the user can quickly identify the purpose, the content and the main access methods of the resource? (cf. Catalogue 5.3)	✓
Acces modes		

Browsing	Does the project offer the possibility to browse the contents by simple browsing options or advanced structured access via indices (e.g. by author, year, genre)? (cf. Catalogue 5.4)	✗
Fulltext search	Does the project offer a fulltext search? (cf. Catalogue 5.4)	✓
Advanced search **Any search that offers more complex search than just a word or a phrase, e.g. boolean operators, wildcards, restricted search, filters or facets.**	Does the project offer an advanced search? (cf. Catalogue 5.4)	✓
Analysis		
Tools	Does the text collection integrate tools for analyses of the data? (cf. Catalogue 5.5)	✓
Customization	Can the user alter the interface in order to affect the outcomes of representation and analysis of the text collection (besides basic search functionalities), e.g. by applying his or her own queries or by choosing analysis parameters? (cf. Catalogue 5.5)	✓
Visualization	Does the text collection provide particular visualizations of the data? (cf. Catalogue 5.6)	no visualization
Personalization	Is there a personalisation mode that enables the users e.g. to create their own sub-collections of the existing text collection? (cf. Catalogue 5.7)	✓
Preservation		
Documentation	Does the text collection provide sufficient documentation about the project in general as well as about the aims, contents and methods of the text collection? (cf. Catalogue 6.1)	✓

Open Access **Are the contents of the presentation freely accessible without subscription fee?**	Is the text collection Open Access? (cf. Catalogue 6.2)	✓
Rights		
Declared	Are the rights to (re)use the content declared? (cf. Catalogue 6.2)	✓
License **CC0: Creative Commons license CC0 applied.** **CC-BY: Creative Commons license CC-BY applied.** **CC-BY-ND: Creative Commons license CC-BY-ND applied.** **CC-BY-NC: Creative Commons license CC-BY-NC applied.** **CC-BY-SA: Creative Commons license CC-BY-SA applied.** **CC-BY-NC-ND: Creative Commons license CC-BY-NC-ND applied.** **CC-BY-NC-SA: Creative Commons license CC-BY-NC-SA applied.** **PDM: Work is in the Public Domain.**	Under what license are the contents released? (cf. Catalogue 6.2)	No explicit license / all rights reserved
Persistent identification and addressing **DOI: Digital Object Identifier according to the definition of The International DOI Foundation. The DOIs should be resolvable through http://doi.org/.** **ARK: Archival Resource Key according to the definition of the California Digital Library. An ARK URL contains the label: 'ark' after the URL's hostname.** **URN: Uniform Resource Name using the urn: scheme. URNs always start with the label 'urn:'.** **PURL.ORG: Persistent Uniform Resource Locator using the PURL concept and administered by the Online Computer Library Centre.** **other service: Choose this if an external service other than the above options is used.** **Persistent URLs: Choose this if the project promises permanent URLs or uses a local resolving system between URLs and underlying technical addresses but does not use any of the external services mentioned in the options.** **none: Choose this if no persistent identifiers and addressing system are used at all.**	Are there persistent identifiers and an addressing system for the text collection and/or parts/objects of it and which mechanism is used to that end? (cf. Catalogue 6.3)	none
Citation	Does the text collection supply citation guidelines? (cf. Catalogue 6.3)	✓

Archiving of the data **Choose yes if you have reason to believe that the archiving and long term sustainability of the data is cared for (e.g. because the data is part of a platform that cares for these aspects), even if the documentation makes no explicit statement about it.**	Does the documentation include information about the long term sustainability of the basic data (archiving of the data)? (cf. Catalogue 6.4)	✗
Institutional curation **Select yes, if there is either an explicit claim that continuous maintenance for the project is provided by some institution or you have strong reason to believe that this is the case, even if it is not explicitly claimed, otherwise select no.**	Does the project provide information about institutional support for the curation and sustainability of the project? (cf. Catalogue 6.4)	✓
Completion **Choose 'yes' if you consider the collection complete. Choose 'no' if further additions and modifications are promised for the text collection to be completed.**	Is the text collection completed? (cf. Catalogue 6.4)	unknown