Aim and concepts

Sketching the world of corpora

DK-CLARIN WP 2.1 Technical Report
Jørg Asmussen, DSL
Final version of May 5, 2015[1]

Deliverables concerned

This report concerns the DK-CLARIN work package 2.1 *Reference corpus of general language* as well as corpus projects carried out by Det Danske Sprog- og Litteraturselskab, dsl.dk. An overview of project-specific deliverables, that is *project tasks*, is given in Section 2.

[1] A more recent version may be available at:
http://korpus.dsl.dk/clarin/corpus-doc/concepts.pdf

Outline

This report gives an overview of DK-CLARIN work package 2.1 *Reference corpus of general language* as a whole and sketches major concepts of this work package and corpus projects carried out by DSL in general.

1 Aim of the project **2**
1.1 Reference corpus 2
1.2 CMRS framework 3
2 Project tasks and documentation outline **5**
3 Text collection, text bank, corpus **10**
3.1 Text collection/archive/repository 10
3.2 Text bank 10
3.3 Corpus 11
4 Document history **12**
5 References **13**

1 Aim of the project

The aim of the project[2] is twofold:

Corpus: Gather a reference corpus of general Danish according to certain design principles.

To achieve this goal, it is necessary to develop a framework for managing the construction process of corpora. This leads to the second aim:

Framework: Establish a *Corpus Management and Retrieval System* (CMRS) as a framework for building and analyzing corpora. The result of this is an in-house (DSL) collection of methods, tools, and means of structured data storage, together with this comprehensive documentation. The CMRS concept established in WP 2.1 aims at being transferable to other corpus management and retrieval scenarios.

1.1 Reference corpus

The reference corpus of general Danish that is gathered as part of this DK-CLARIN project (WP 2.1) comprises 45 million running tokens and is textually mixed, although – as a matter of rather limited resources – texts from periodicals like newspapers and magazines are preferred as they are more straightforward to process

[2] Throughout the documentation *the project* refers to DK-CLARIN work package 2.1, i.e. building a reference corpus of general language.

especially if they come in some kind of XML format as is the case for texts from one of DSL's main sources, Infomedia.[3]

Metadata The texts of the corpus are provided with metadata given as XML-formatted headers. These ensure that texts can be filtered according to different textual parameters. However, the level of detail is restricted to the general information attached to the material when it is received from the text supplier.

Markup The tokens of the material are tagged with information on lemma forms, part-of-speech, and inflectional markers.

Accessibility A copy of the corpus is included in the DK-CLARIN repository of resources and tools. It is made publicly accessible through a web-based concordancer as well.[4]

1.2 CMRS framework

As the *Corpus Management and Retrieval System* is an essential prerequisite for the accomplishment of the project – working as its "corpus factory" –, we will start with an overview of this system and give a brief description of its components.

Figure .1 shows the *compositional structure* or "architecture" of the CMRS. The figure follows largely the Fundamental Modeling Concepts framework, FMC. However, some minor modifications have been made to a few symbols of the notation framework, in particular, the symbol for read/write access to a storage (two rounded unidirectional arrows in FMC) has been replaced by a strong white bidirectional arrow that consumes less space and can be applied more flexibly. FMC's channel concept (usually denoted by a small circle) has been replaced by a somewhat broader concept of interfaces through which communication between compound components working together as a system and other systems or users outside take place. Figure .2 shows the types of interfaces used in this notation.[5]

As illustrated in Figure .1, the major data repository of the CMRS is the *text bank* that holds four different types of data: a collection of the *texts* themselves, a collection of *metadata* describing the general characteristics of each individual text, a collection of *annotations* providing various linguistic information on text token level, and finally *supplier details* providing information on contacts, text deliverances, and agreements. The text bank is fed with textual material from text suppliers through *import handlers*, that is, transducers that convert the various text formats into the specific one(s) needed in the CMRS. Texts may be manually complemented with metadata through a *metadata editor*, the same goes for supplier details. Annotating the text material with linguistic information is performed

[3] http://infomedia.dk/

[4] The corpus will be included as a stand-alone corpus in KorpusDK and should be publicly available under ordnet.dk/korpusdk from winter 2013-14.

[5] See www.fmc-modeling.org for a quick introduction and pointers to further reading.

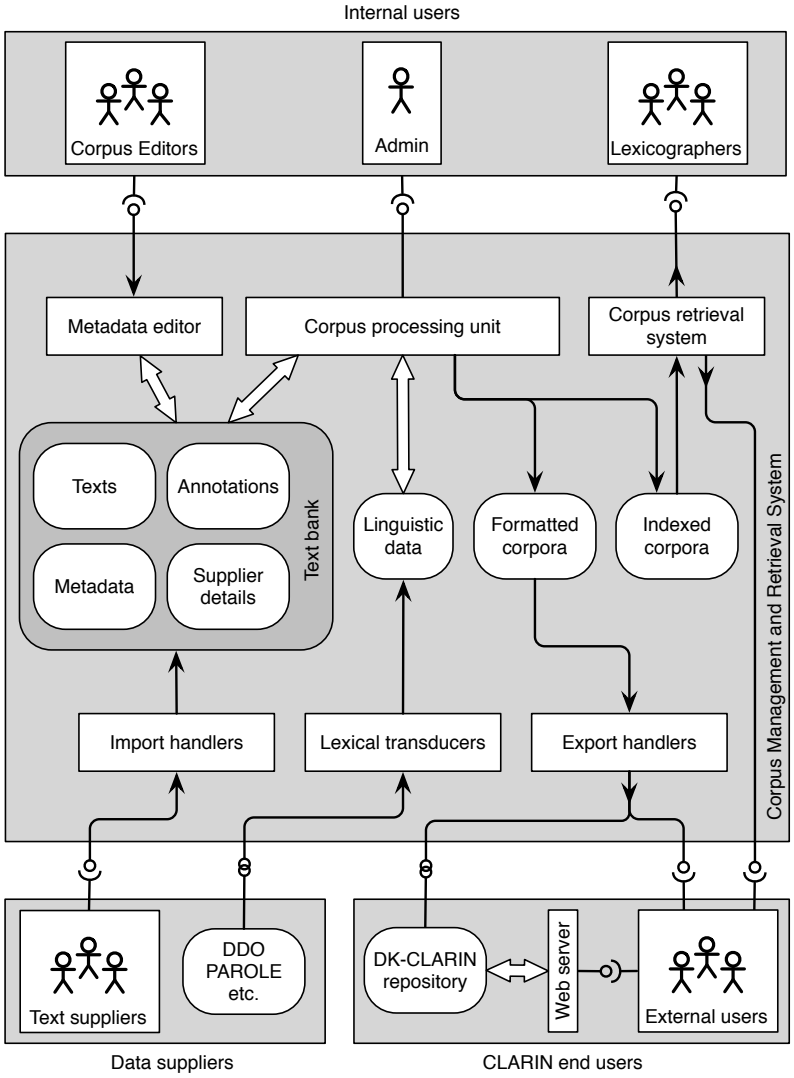

Figure .1: Major components and interfaces of the CMRS

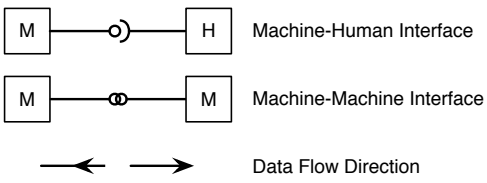

Figure .2: Symbols for various types of interfaces

by the *corpus processing unit* that utilizes linguistic data drawn from external resources and processed by *lexical transducers*. The corpus processing unit also extracts *formatted corpora* according to different structural specifications that may be exported through export handlers. Finally, the corpus processing unit produces *indexed corpora* that can be deployed to external/internal *users* through a *corpus retrieval system* – a corpus search engine like a concordancer for example. The difference between formatted and indexed corpora is that formatted corpora are formatted in accordance with a (TEI) schema of some kind and contain text samples possibly together with metadata and some sort of linguistic annotations on token level. In contrast, indexed corpora are made searchable in some sort of search engine. Metadata and annotations may be searchable as well as may be additional lexical data derived from the material, e.g. different types of frequency lists, collocations, or other statistical material.

2 Project tasks and documentation outline

In order to achieve the major aims of the project, that is building a CMRS and use it for gathering a reference corpus, a number of steps to perform were defined in the original project plan[6] where the result of each of these steps is termed *deliverable*. Some of these steps are directly related to the design and construction of the corpus, others are mainly related to establishing the CMRS. So the list of steps to be taken constitutes an unordered to-do list rather than reflecting the sequential process of either building a CMRS or applying it to "assemble" corpora. As opposed to the list of deliverables, the present documentation is structured in order to reflect the *sequential process* of making a corpus from the *design* phase, over *collecting* and *markup*, to final *deployment*.

In the following, the steps/deliverables of the original plan are listed together with references to the technical reports where the corresponding documentation – that serve as *deliverable reports* – can be found. The documentation proper gives further pointers to the tools and resources.[7]

D1 Text registry DSL as well as DSN collect Infomedia text material, parts of which are likely to be included in the WP 2.1 corpus. Therefore, a way of registering texts needs to be established. A registry allows tracing and eliminating possible duplicate texts. The text registry functionality is part of the CMRS. **Outcome:** Report.

→ Asmussen (2013g)

[6] http://korpus.dsl.dk/clarin/wp21/wp21-arbejdsplan-old.pdf

[7] The CMRS itself is not considered a DK-CLARIN deliverable as it is not possible to fully implement it due to limited resources. Instead accounts of its different components is given in this series of technical reports. Based on the documentation, it should be straightforward for other, similar projects to reuse some of the design considerations applied.

D2 Tokenizer A consistent and easy-to-use token concept needs to be defined. The token concept has important implications on the design of the tokenizer tool and the POS-tagger applied in WP 2.1. **Outcome:** Tool and report.

→ Asmussen (2013d)

→ Asmussen (2013e)

D3 Decision on text bank system A text bank system is necessary for project-internal text administration. Investigations of different approaches to such a system will be carried out. Two general options seem viable – either one based on a relational db or on an XML db. The text bank system is the core component of the CMRS. **Outcome:** Report.

→ Asmussen (2013g)

D4 Text supplier registry A registry of active and potential text suppliers needs to be designed as an integrated component of the CMRS. **Outcome:** Report.

→ Asmussen (2013g)

D5 Implementation of text bank system The chosen text bank approach (see D3) implemented (possibly with a GUI) as component of the CMRS. **Outcome:** Report and a project-internal service.

→ Asmussen (2013g)

D6 Processing of Infomedia text Conversion to the DK-CLARIN format and text bank import of Infomedia material collected by DSL and DSN in 2008 and 2009. **Outcome:** Report.

→ Asmussen (2013e)

D7 Development of format transducers Design and development of transducers capable of transforming all supplier formats into the WP2.1 text format. **Outcome:** Report and project-internal services.

→ Asmussen (2013e)

D8 Processing of other text Collected text material is converted and inserted into the text bank component of the CMRS. **Outcome:** Report.

→ Asmussen (2013e)

D9 Full-form lexicon Development and/or configuration of a full-form lexicon for POS tagging. **Outcome:** Resource with documentation.

→ Asmussen (2013f)

D10 Lemmatizer It is considered indispensable that corpus texts need to indicate the lemma form of each inflected word form in the corpus to let the user of the corpus perform more flexible queries. Therefore, it is necessary to either develop or configure a lemmatizer (that may be based on a full-form lexicon or a morphological analyzer). In the context of WP 2.1, a lemmatizer designed as an integral part of a POS tagger is the preferable solution. **Outcome:** Tool with documentation.

→ Asmussen (2013c)

→ Asmussen (2014)

→ Asmussen (2013f)

D11 POS tagger In order to tag tokens in corpus texts with part-of-speech information, it is necessary to either develop or configure a POS tagger (either based on a full-form lexicon or a morphological analyzer) and a suitable tag set. **Outcome:** Tool with documentation.

→ Asmussen (2013c)

→ Asmussen (2014)

→ Asmussen (2013f)

D12 Download service Implementation of download option for copyright-cleared or scrambled text material. **Outcome:** None. Cancelled due to project cutbacks.

D13 TEI transducer The original plan for WP 2.1 was based on the assumption that the repository of potential corpus texts – the corpus text bank – most likely would have a non-XML structure (relational db). In order to make interchange of texts easy and in order to make them fit into the intended resource repository of DK-CLARIN, the development of a transducer that could reshape the texts and metadata stored in the corpus text bank to valid TEI XML seemed necessary. However, during the course of the project, it became clear that the text bank itself should be implemented as an XML database so that the texts could be stored in their final TEI XML format. Therefore, the task of developing a transducer became a task of defining an appropriate subset of TEI in order to suit the metadata and text format needs of DK-CLARIN. **Outcome:** Report.

→ Asmussen (2015)

→ Asmussen (2013d)

→ Asmussen (2013e)

D14 Prototype of concordance tool A web-based concordance tool needs to be configured/implemented as a prototype for testing. **Outcome:** Report.

→ Asmussen (2013a)

D15 Panel of test users Constitution of a panel of test users. **Outcome:** None. Cancelled due to project cutbacks.

D16 User tests Performing and evaluating user tests of web-concordancer. **Outcome:** None. Cancelled due to project cutbacks.

D17 Final version of concordance tool Web-based concordancer with public access. **Outcome:** Service with documentation.

→ Asmussen (2013a)

D18 Final version of corpus Final version of POS-tagged corpus of 45 million words available for the DK-CLARIN repository and accessible through a web-based (or other) concordance tool. **Outcome:** Resource with documentation.

→ Asmussen (2013b)

As already mentioned above, the reports do not follow the order of steps/deliverables but instead a more general structure as the process of building a corpus can be subdivided into four phases:

1. **Design:** Textual metadata must be determined as well as annotations on other textual levels. A repository for storing the text material – a text bank – needs to be designed and implemented as part of the CMRS. Designing a text bank includes designing the representation of text data.

   The following technical reports cover the design phase of the project:

   → Text metadata: Asmussen (2015)

   → Text formatting: Asmussen (2013d)

   → Text bank: Asmussen (2013g)

   These chapters cover the following tasks/deliverables of the project:

   → D 1 – D 5 and D 13

2. **Collecting:** This phase covers negotiations with potential text suppliers, the conversion of text material gathered in a myriad of odd formats into the standard format defined during the design phase, and finally storing it in the text bank. It also covers the task of manually adding missing metadata.

   The following technical report covers the phase of collecting material:

   → Text processing: Asmussen (2013e)

   A technical report on text acquisition would be desirable as well, however, as this documentation focuses on the technical aspects of building a corpus, it has been left out.

   This chapter covers the following tasks/deliverables of the project:

   → D 2, D 6 – D 8, and D 13

3. **Markup:** Texts that have been converted to the standard format are ready to have various types of annotations added. In this project, all words in a text are tagged with the following types of annotations: A standardized orthographic form of the word, its lemma form, its part of speech as well as some inflectional information. In order to carry out these types of markup, certain tools need to be developed and/or configured as part of the markup phase.

   The following technical reports cover the markup phase:

   → Survey of POS taggers: Asmussen (2013c)

   → Design of the jaPOS tagger: Asmussen (2014)

   → The full-form lexicon: Asmussen (2013f)

   These chapters cover the following tasks/deliverables of the project:

   → D 9 – D 11

4. **Deployment:** Deploying the corpus means to make it accessible for end-users, either through a corpus retrieval system of some kind that needs to be developed/configured, or by distributing the text files, that make up the corpus, in a standard format.

   The following technical reports cover the markup phase:

   → Corpus specifications: Asmussen (2013b)

   → Forthcoming *PAROLE version 2* documentation

   → Forthcoming *Corpus retrieval software (CoREST)* documentation

   → Corpus access: Asmussen and Offersgaard (2008)

   This chapter covers the following tasks/deliverables of the project:

   → D 14 and D 17

3 Text collection, text bank, corpus

This Section aims at giving some clarification on the concepts *text collection*, *text bank*, and *corpus*, of which the two latter play an important role in the CMRS.

3.1 Text collection/archive/repository

A text collection is a collection of complete or abridged texts of any kind collected by a project or institution/company. The purpose of collecting the material may be documentation or archiving in general for purposes not yet defined – and often, corpus construction is not considered an option at all. As a consequence of not having specified an explicit purpose for the text collection other than maybe documentation, the process of collecting is often opportunistic rather than guided by certain corpus-compositional principles. The texts of a text collection may carry a well-defined minimum of annotations on text level. Based on these annotations, a subset of text items may be exported from the the archive. An archive may be a structured means of storage, e.g. a database holding the texts in some generalized format.[8] However, there does not need to be any structured means of storage – the material may reside unordered in a file system and may be composed of texts with various incompatible formats. In relation to the CMRS, it may be considered text material that has not yet been imported into the CMRS but is stored in other accessible locations and formats.

3.2 Text bank

Whereas a text collection still may be rather unstructured, a text bank is an implementation geared to storing and retrieving texts to be potentially included in linguistic corpora. So, in contrast to a text collection, the explicit purpose of the texts gathered in a text bank is to be able to build corpora from them. It allows to better process and organize potential corpus text material.[9] Therefore, a text bank requires an elaborate structure: Texts of a text bank must carry well-defined meta-information (e.g. expressed as an XML-formatted header), they must follow a well-defined format, that is, they should be tokenized and each token should have a unique reference ID in order to be addressed unambiguously. A text bank provides furthermore interfaces/handlers through which texts and metadata can be added and through which texts can be selected for export as a corpus in an appropriate format. Corpus query functions for linguistic investigation are not part of the text bank, but are features of separate corpus query and statistics tools. The text bank is first and foremost a tool for text and corpus administration.

[8] Examples of text collections are The Oxford Text Archive and Arkiv for Dansk Litteratur.

[9] The text bank must not be confused with the general DK-CLARIN repository developed in WP 5 that is supposed to support various data types (e.g. texts, images, lexicons) and various formats to be used in various contexts, not just corpus construction.

3.3 Corpus

A corpus is a group of text items from a text bank that have been selected due to explicit criteria based on the information given in the meta-data part of a text item. The purpose of a corpus is to allow certain linguistic investigations as it is assumed that the corpus (text items as whole) constitutes a representative sample of the sort of language to be investigated. Each text item in a corpus either carries the same meta-data that are used in the text bank or a subset of them, e.g. only those that are relevant for the corpus in question. A corpus can be supplemented with meta-data describing its characteristics, e.g. the purpose of it and the selection criteria for the texts it is comprised of. Corpus texts usually carry several token annotation layers, e.g. an orthographically normalized version of the token, a lemmatized form of it, POS information, inflectional information. A corpus of this type can be made accessible for queries in a concordancer or it may be used to be processed by other corpus tools, e.g. for statistical purposes.

4 Document history

A more recent version of this report may be downloaded here:
http://korpus.dsl.dk/clarin/corpus-doc/text-format.pdf

Current version (May 5, 2015)

→ Finalized.

5 References

Asmussen, J. (2013a). Corpus access. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/access.pdf.

Asmussen, J. (2013b). Specifications of the DK-CLARIN reference corpus. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf.

Asmussen, J. (2013c). Survey of POS taggers. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-survey.pdf.

Asmussen, J. (2013d). Text formatting. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-format.pdf.

Asmussen, J. (2013e). Text processing. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-processing.pdf.

Asmussen, J. (2013f). The full-form lexicon. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf.

Asmussen, J. (2013g). The text bank. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/textbank.pdf.

Asmussen, J. (2014). Design of the ePOS tagger. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/pos-design.pdf.

Asmussen, J. (2015). Text metadata. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/text-header.pdf.

Asmussen, J. and Offersgaard, L. (2008). Working with corpora – some scenarios. Technical report, DK-CLARIN, korpus.dsl.dk/clarin/corpus-doc/scenarios.pdf.