ride. A review journal for digital editions and resources

published by the IDE

Rezension von „Europarl“

Europarl, Philipp Koehn (ed.), 2001-2012. http://www.statmt.org/europarl/ (Last Accessed: 19.01.2018). Reviewed by Claes Neuefeind (Institute for Digital Humanities, University of Cologne), c.neuefeind (at) uni-koeln.de.

Abstract

The *Europarl* corpus, short for „European Parliament Proceedings Parallel Corpus 1996-2011“, is provided by the School of Informatics, University of Edinburgh. *Europarl* was first published in 2001, the latest release (May 15th, 2012) includes the parliament proceedings from 1996-2011. There is no dedicated user interface, since *Europarl* is primarily meant to support research on machine translation, which has to rely on parallel texts. Beyond that, it is well suited for other NLP research involving multilingual issues. The resource is freely downloadable. Being one of the standard resources in the field of statistical machine translation, it comes as a source release with accompaining tools that were used for compiling the resource. The simple annotation and formatting makes the source files easy to use in different contexts of application.

Einleitung

1 Das *Europarl*-Korpus (http://www.statmt.org/europarl[1]) entstand an der University of Edinburgh unter der Leitung von Philipp Koehn. *Europarl* steht für „European Parliament Proceedings Parallel Corpus“, womit die wesentlichen Merkmale der Ressource bereits benannt sind: Es handelt sich um ein Parallelkorpus, das Texte in verschiedenen Sprachversionen bereithält. Grundlage des Korpus bilden die Sitzungsprotokolle des Europäischen Parlaments, die regelmäßig auf dessen offizieller

Website veröffentlicht werden.[2] Das Korpus macht sich eine Besonderheit dieser Sitzungsberichte zu Nutze, namentlich deren Mehrsprachigkeit. Diese ist fest verankert in den Europäischen Verträgen, die eine vollständige Gleichbehandlung aller Amtssprachen der Europäischen Union (EU) verlangen. Auf dieser Grundlage werden alle wichtigen EU-Dokumente, einschließlich der parlamentarischen Texte, in den 24 offiziellen Amtssprachen der EU veröffentlicht, d.h. es existieren zu jedem Text, der veröffentlicht wird, stets 23 parallel übersetzte Versionen.[3]

Eckdaten

2 Das *Europarl*-Korpus wurde erstmals im Jahre 2001 veröffentlicht. Seither gab es eine Reihe von Aktualisierungen, die immer wieder auch kleinere Format-Anpassungen beinhalten. Die Arbeiten am *Europarl*-Korpus wurden u.a. im Rahmen des *EuroMatrixPlus*-Projekts im 7. Rahmenprogramm der Europäischen Kommission gefördert.[4] Die bis heute letzte Version wurde kurz nach dem Ende der Projektförderung am 15.5.2012 veröffentlicht und umfasst die Sitzungsprotokolle des Europäischen Parlaments der Jahre 1996 bis 2011. Gegenüber der ersten Version, die zunächst nur Texte in 11 Sprachen aus dem Zeitraum von April 1996 bis Dezember 2001 umfasste, mit durchschnittlich ca. 20 Millionen Wörtern bzw. ca. 740.000 Sätzen je Sprache, ist das Korpus mit jeder Version weiter angewachsen. Die aktuelle Version *v7* umfasst insgesamt 753 Millionen Wörter bzw. 30 Millionen Sätze, die sich auf Texte in 21 Sprachen mit bis zu 55 Millionen Wörtern bzw. mehr als 2 Millionen Sätzen verteilen, wobei die Zahlen für die einzelnen Sprachen zum Teil erheblich voneinander abweichen.[5]

Eigenschaften

3 Die charakteristischen Eigenschaften der Ressource ergeben sich unmittelbar aus dem spezifischen Anwendungsfall, für den das *Europarl*-Korpus entworfen wurde: Parallelkorpora sind Grundlage und Voraussetzung für die Entwicklung von Systemen zur statistischen maschinellen Übersetzung (*Statistical Machine Translation*, SMT). Dementsprechend bestand die wesentliche Motivation für die Zusammenstellung des *Europarl*-Korpus darin, Daten für die Entwicklung und den Test von SMT-Systemen bereitzustellen. SMT galt von den 1990er Jahren bis in die Mitte der 2010er Jahre hinein als *state-of-the-art* im Bereich der Maschinellen Übersetzung. Auch wenn neuere Ansätze heute verstärkt auf den Einsatz von Künstlichen Neuronalen Netzen setzen,[6]

so ist SMT dennoch nach wie vor die am weitesten verbreitete Methode zur Maschinellen Übersetzung. SMT-Systeme basieren auf statistischen Lernverfahren, die auf ausreichend große Mengen von Trainings- und Testdaten angewiesen sind, um robuste Ergebnisse zu erzielen und um diese angemessen evaluieren zu können. Als Trainingsdaten dienen dabei Paare von (Teil-)Sätzen bzw. Phrasen (je nach konkretem Ansatz), die über längere Textabschnitte hinweg aligniert, d.h. einander parallel zugeordnet werden.

4 Nachdem in den späten 1990er Jahren kaum Ressourcen von ausreichender Größe vorlagen, sollte die Erstellung des *Europarl*-Korpus diese Lücke schließen. Als Ausgangsmaterial für ein Parallelkorpus sind die Sitzungsberichte des Europäischen Parlaments in besonderer Weise geeignet, da sie eine der größten frei verfügbaren Quellen für die Akquise parallel übersetzter Texte darstellen. Die Auswahl der *Europarl*-Daten hat dabei gleich eine ganze Reihe positiver Nebeneffekte: Weil die Daten direkt vom Europäischen Parlament herausgegeben werden, ist zum einen ein Mindestmaß an Qualitätssicherung in Bezug auf die Zuverlässigkeit der Übersetzungen gegeben, zum anderen ist damit auch die Copyright-Frage geklärt, da die Texte per Gesetz frei zugänglich sein müssen. Darüber hinaus ist durch eine entsprechende Selbstverpflichtung des Europäischen Parlaments auch die Langzeitarchivierung der Ausgangsdaten gesichert. Solange die zur Erstellung des Europarl-Korpus eingesetzten Tools zur Verfügung stehen, können damit nicht nur die über die Webseite archivierten Releases im Grunde jederzeit reproduziert werden, sondern es können theoretisch auch eigene, bspw. um neuere Daten erweiterte Versionen des Korpus erstellt werden.

Design und Methodik

5 Bei der Erstellung des Korpus standen in erster Linie die Einfachheit und Kontrollierbarkeit des Verfahrens im Vordergrund. Das in Koehns (2002, 2005) beschriebene Vorgehen gliedert sich im Wesentlichen in die Schritte Datenakquise, Alignierung der Texte sowie die Zerlegung in einzelne Sätze und deren parallele Zuordnung. In einem ersten Schritt werden die Texte mithilfe eines Crawlers in Form von HTML-Dateien von der offiziellen Website des Europäischen Parlaments bezogen. Vorteil ist neben den oben genannten Aspekten (Qualitätskontrolle, Copyright, Langzeitarchivierung) auch die Tatsache, dass die Daten bereits in einem weitgehend homogenen Format vorliegen. Anschließend werden aus den HTML-Dateien Informationen über die jeweiligen Sprecher, das Thema (bzw. den thematischen *thread*,

in dem ein Redebeitrag steht), die verwendete Sprache und das Datum extrahiert. Die Dokumente werden anhand der enthaltenen Zwischenüberschriften automatisiert in thematische Abschnitte gegliedert und durch die Vergabe fortlaufender Ids über die verschiedenen Sprachversionen hinweg aligniert. In einem weiteren Schritt werden die Texte unter Verwendung eines zusammen mit dem Korpus bereitgestellten Perl-Skriptes automatisiert in einzelne Sätze zerlegt, die abschließend mit einem weiteren Skript auf Basis einer Variante des Algorithmus von Gale und Church (1993) aligniert werden.[7]

6 Aus Perspektive der Digital Humanities durchaus kritisch zu bewerten ist die Tatsache, dass im *Europarl*-Korpus teilautomatisiert erstellte Trainingsdaten als Basis für die Entwicklung eines computerlinguistischen Verfahrens eingesetzt werden. Jedoch ist dieses Vorgehen in der Computerlinguistik durchaus üblich, insbesondere im Bereich der SMT, bei der eine sehr umfangreiche Datengrundlage Voraussetzung ist. Um die Fehler bei Tokenisierung und Satztrennung zu minimieren, werden zum einen sprachspezifische Anpassungen vorgenommen, z.B. durch Abkürzungsverzeichnisse. Zum anderen wird durch die vorherige Unterteilung der Dokumente in kurze Absätze von ca. 2-5 Sätzen verhindert, dass sich evtl. auftretende Alignment-Fehler über Absatzgrenzen hinaus fortsetzen können. Fehlerfreiheit kann zwar auch durch diese Maßnahmen nicht vollständig garantiert werden, jedoch können sie aufgrund der Robustheit der eingesetzten statistischen Verfahren in der Regel vernachlässigt werden.

Korpusstruktur

7 Resultat der verschiedenen Verarbeitungsschritte sind zunächst die alignierten Dokumente, die jeweils in einer eigenen Datei je Sitzungsdatum und Sprache gespeichert werden. In diesen Dateien sind nur rudimentäre, für das paarweise Alignment von (Teil-)Abschnitten benötigte Annotationen enthalten. Wie in Code 1 dargestellt, beschränken sich diese auf die Markierung des Themas, in dem ein Redebeitrag steht (kodiert als <CHAPTER id>, welche demnach nicht Kapitel, sondern vielmehr thematische Abschnitte bezeichnen), den jeweiligen Sprecher und die Sprache (<SPEAKER id name language>, wobei die Sprachangabe optional ist), sowie die Markierung einzelner Unterabschnitte (<P>).

```
<p><CHAPTER ID=1> Resumption of the session <SPEAKER ID=1 NAME="President"> I
declare resumed the session of the European Parliament ... <P> Although, as
you will have seen, the dreaded 'millennium bug' ...
```

Code 1: Annotationen im Europarl-Korpus, hier am Beispiel der ersten Zeilen der Datei txt/en/ep-00-01-17.txt des source release in der Version v7, welche das entsprechend aufbereitete Sitzungsprotokoll vom 17. Januar 2000 enthält (https://web.archive.org/web/20170619073042/http://www.statmt.org/europarl/v7/europarl.tgz).

8 Ein weiteres Ergebnis des Erstellungsprozesses sind satzweise alignierte Parallelkorpora, in denen jeder Satz in einer eigenen Zeile steht, so dass die korrespondierenden Sprachversionen anhand der Zeilennummer identifiziert werden können. Im Gegensatz zu den Dokument-Dateien, die als Ausgangsmaterial dienen, enthalten die Parallelkorpora nurmehr die rohen Textdaten, d.h. das minimale Markup der Ausgangsdateien wird vollständig entfernt. Das *Europarl*-Korpus wird als sog. *source release* bereitgestellt, welcher zum einen die alignierten Dokumente für alle 21 berücksichtigten Sprachen enthält, zum anderen auch die für deren Aufbereitung eingesetzten Tools zur satzweisen Alignierung, so dass gewissermaßen *on demand* Parallelkorpora für verschiedene Sprachpaare erstellt werden können (potentiell bis zu 20 für jede Sprache). Bei den im *source release* enthaltenen Tools handelt es sich um eine Reihe von einfachen Perl-Skripten für die Korpusaufbereitung, für deren Ausführung eine Perl-Installation vorhanden sein muss. Diese umfassen ein Skript für die satzweise Alignierung, ein Skript für die Konversion in XML-Dateien, einen *Sentence-Splitter* und einen *Tokenizer*, die durch ein umfassendes, sprachspezifisch angelegtes Abkürzungsverzeichnis ergänzt werden, das all jene Fälle enthält, in denen Punkte als Bestandteil des Wortes erkannt werden sollen. Zusätzlich zu diesem *source release* stehen über die Website 20 Parallelkorpora, jeweils ausgehend vom Englischen, in vorbereiteter Form zur Verfügung und können dort einzeln heruntergeladen werden.

Einsatzmöglichkeiten

9 Haupteinsatzgebiet für das *Europarl*-Korpus sind sprachtechnologische Anwendungen, im Besonderen im Kontext der Entwicklung von SMT-Systemen. Über die Arbeiten von Philipp Koehn bzw. der Edinburgh-Gruppe hinaus wird das *Europarl*-Korpus u.a. seit 2005 regelmäßig im Rahmen sogenannter *Shared Tasks* bei Konferenzen der *Association for Computational Linguistics* (ACL) für die kompetitive Entwicklung und Evaluation von SMT-Systemen genutzt.[8] Die Einsatzmöglichkeiten gehen jedoch über die maschinelle Übersetzung hinaus. So können die *Europarl*-Daten

im Grunde für alle Arten von NLP-Tasks eingesetzt werden, die multilinguale Daten erfordern; Koehn (2005) verweist hier etwa auf Arbeiten im Bereich der Wortsinndisambiguierung, der Anaphernresolution oder auch der Informationsextraktion.[9]

10 Das *Europarl*-Korpus ist somit primär als sprachtechnologische Ressource zu verstehen und ist demnach nicht als Datenbasis für die Recherche oder für korpuslinguistische Untersuchungen konzipiert. Für die Nutzung der Daten in anderen Anwendungskontexten gibt es jedoch eine Reihe von Alternativen. So findet sich bspw. auf der Webseite des OPUS-Projekts des *Nordic Language Processing Laboratory* (NLPL),[10] das eine Sammlung frei verfügbarer Parallelkorpora bereitstellt, ein Suchinterface für eine auf Basis der *Open Corpus Workbench* (CWB)[11] linguistisch aufbereitete Version des Europarl-Korpus. Eine vergleichbare Zugriffsmöglichkeit bietet auch die (allerdings kostenpflichtige) Korpusmanagement-Software SketchEngine (vgl. Kilgarriff et al. 2014), in der eine Vielzahl von verschiedenen Korpora gebündelt sind, darunter auch eine größere Anzahl frei verfügbarer Ressourcen wie eben das *Europarl*-Korpus.[12]

11 Weil die zugrunde gelegten Daten aufgrund ihres Inhalts insbesondere auch für den Bereich der Politikwissenschaft von großem Interesse sind, wurde 2014 im Rahmen des Projekts *Talk of Europe* mit dem *LinkedEP Dataset*[13] eine alternative Version zu *Europarl* veröffentlicht, die auf Semantic-Web-Technologien basiert (vgl. van Aggelen et al., 2016). Während das *Europarl*-Korpus dezidiert für den Anwendungsfall der maschinellen Übersetzung entworfen wurde, zielt das *LinkedEP Dataset* auf eine inhaltlich orientierte Durchsuchbarkeit der Daten, indem diese mit zusätzlichen Informationen angereichert werden und so eine Kontextualisierung der enthaltenen Informationen ermöglichen. Es handelt sich um eine nach Linked Open Data (LOD)-Prinzipien kodierte Ressource, die wie das ursprüngliche *Europarl*-Korpus von den auf der offiziellen Website des Europäischen Parlaments als HTML-Seiten veröffentlichten Sitzungsprotokollen ausgeht und diese in RDF überführt. Die resultierende Ressource ist zudem mit weiteren LOD-Ressourcen verknüpft, u.a. *GeoNames*[14], *DBPedia*[15], sowie der *Automated Database of the European Parliament*[16], die bibliographische Daten der Mitglieder des Europäischen Parlaments bereitstellt.

Abschließende Bemerkungen

12 Die Besonderheit des *Europarl*-Korpus gegenüber anderen Parallelkorpora, wie sie bspw. über das o.g. OPUS-Projekt bereitgestellt werden, ist neben der hohen Anzahl der möglichen Sprachpaare und dem daraus resultierenden Umfang vor allem die verbürgte Qualität der Übersetzungen sowie die Tatsache, dass aufgrund ihrer Herkunft aus dem Umfeld des Europäischen Parlaments die oftmals problematischen Fragen bezüglich des Copyrights sowie der langfristigen Verfügbarkeit der Daten geregelt sind.

13 Das *Europarl*-Korpus ist zudem eine sehr stark spezialisierte Ressource, die als multilinguales Parallelkorpus dezidiert für die Entwicklung von Systemen zur statistischen maschinellen Übersetzung (*Statistical Machine Translation*, SMT) angelegt wurde. In ihrer Beschränkung auf einen einzelnen Anwendungsfall stellt das Korpus gegenüber anderen Sprachressourcen sicherlich einen Sonderfall dar, sowohl innerhalb der (Computer-)Linguistik als auch im erweiterten Bereich der Digital Humanities (DH). Insbesondere der sehr sparsame Einsatz von Annotationen mag zunächst sehr minimalistisch anmuten, jedoch hat der geringe Grad an Modellierung der Daten hier durchaus Methode. So ist der weitgehende Verzicht auf Annotationen vor allem dem sprachtechnologischen Anwendungsfall geschuldet: Da in SMT-Systemen allein die alignierten Sprachpaare betrachtet werden, wird jede zusätzliche Information bewusst aus der Verarbeitung herausgehalten. Während andere, v.a. korpuslinguistische Ressourcen in der Regel auf verschiedenen Ebenen mit Annotationen angereichert sind und dadurch differenzierte linguistische Analysen ermöglichen, zeichnet sich das *Europarl*-Korpus durch den bewussten Verzicht auf solche Auszeichnungen aus. Wenngleich es sich um ein älteres Korpus handelt, so entspricht diese Art der Aufbereitung nach wie vor den üblichen Standards zur Erstellung von Parallelkorpora für die maschinelle Übersetzung. Das *Europarl*-Korpus wird dementsprechend noch immer häufig eingesetzt, etwa als *baseline* bzw. *benchmark* in der Entwicklung und Evaluation klassischer SMT-Systeme, aber auch im Kontext neuerer Ansätze auf Grundlage von Künstlichen Neuronalen Netzen (sog. „Neural Machine Translation“, NMT),[17] die für die Erstellung von Sprachmodellen u.a. auch alignierte Parallelkorpora als Input nutzen.[18]

14 Für Anwendungen im Bereich der DH ist das *Europarl*-Korpus in dieser rohen Form dagegen nur bedingt nachnutzbar. Neben der weitgehend fehlenden Datenmodellierung liegt dies nicht zuletzt auch am Fehlen eines Suchzugriffs zu Recherchezwecken. Zwar stehen hier mit dem o.g. OPUS-Suchinterface sowie der

Korpusmanagement-Software *SketchEngine* entsprechende Alternativen zur Verfügung, die eine korpuslinguistische Nutzung des Korpus ermöglichen; aus Perspektive der DH ist jedoch sicherlich vor allem die stärker inhaltlich motivierte Erschließung der Daten durch die Semantic-Web-Ressource *LinkedEP* von größerem Interesse, da diese deutlich über den Anwendungsfall der SMT hinausweist, indem sie eine Kontextualisierung des enthaltenen Wissens ermöglicht. Nachdem der hierfür nötige Grad an zusätzlicher Datenmodellierung weit über das ursprüngliche *Europarl*-Korpus hinausgeht, handelt es sich hier jedoch im Grunde um eine unabhängige Ressource, die zwar von der gleichen Datengrundlage ausgeht, ihrerseits aber ebenfalls nur für einen spezifischen Anwendungsbereich angelegt ist – eine Aufbereitung der Texte, die für beide Anwendungsbereiche funktionieren würde, erscheint vor diesem Hintergrund nicht wirklich praktikabel.

15 Vielmehr ist in der Beschränkung auf den Bereich sprachtechnologischer Anwendungsfälle einer der wesentlichen Gründe für die Bekanntheit und Verbreitung des *Europarl*-Korpus zu sehen. Besonders hervorzuheben ist hierbei das Prinzip des *source release*, bei dem die nur minimal aufbereiteten Ausgangsdaten zusammen mit den Tools für die satzweise Alignierung veröffentlicht werden, so dass nicht nur die Quellen rekonstruiert werden können, sondern auch die Erstellung von Parallelkorpora für einzelne Sprachpaare nach Bedarf vorgenommen werden kann. Wenngleich also die Präsentation der Daten sehr minimalistisch ausfällt und die Website im Vergleich zu anderen korpuslinguistischen Ressourcen vielleicht etwas anachronistisch anmuten mag, so ist das *Europarl*-Korpus dennoch für den Bereich der maschinellen Übersetzung als ein *best practice* für die Erstellung von Parallelkorpora anzusehen, und wird als solches sicherlich auch weiterhin eine große Rolle spielen.

Anmerkungen

1. Archivierter Link: https://web.archive.org/web/20171228005236/http://www.statmt.org/europarl.

2. https://web.archive.org/web/20171224050712/http://www.europarl.europa.eu/.

3. Siehe dazu auch die Geschäftsordnung des Europäischen Parlaments: https://web.archive.org/web/20171002224154/http://www.europarl.europa.eu/aboutparliament/de/20150201PVL00013/Mehrsprachigkeit.

4. https://web.archive.org/web/20170629163336/http://www.euromatrixplus.net.

5. Für eine detaillierte tabellarische Aufstellung siehe https://web.archive.org/web/20171228005236/http://www.statmt.org/europarl. Über die Webseite können zudem weiterhin auch die älteren Versionen bezogen werden.

6. Als wohl bekanntestes Beispiel sei hier der Übersetzungsservice Google Translate genannt, der seit November 2016 schrittweise von SMT auf einen Deep-Learning-Ansatz umgestellt wird (siehe https://web.archive.org/web/20180104154107/https://www.blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/).

7. Zu Verfügbarkeit und Umfang der mit dem Korpus bereitgestellten Verarbeitungswerkzeuge siehe den nachfolgenden Abschnitt.

8. Eine Aufstellung der entsprechenden ACL-Workshops findet sich unter http://www.statmt.org.

9. Speziell für sprachtechnologische Anwendungen wurde an der Universität Zürich auch eine bereinigte und mit zusätzlichen Strukturinformationen angereicherte Version des *Europarl*-Korpus erstellt; siehe dazu Graën et al. (2014) sowie https://web.archive.org/web/20141225132930/http://pub.cl.uzh.ch/purl/costep.

10. https://web.archive.org/web/20180105145759/http://opus.nlpl.eu/.

11. https://web.archive.org/web/20170915061225/http://cwb.sourceforge.net/.

12. https://web.archive.org/web/20171109044747/https://www.sketchengine.co.uk/.

13. https://web.archive.org/web/20161230090013/http://www.talkofeurope.eu/data.

14. https://web.archive.org/web/20171228152250/http://www.geonames.org.

15. https://web.archive.org/web/20170810113013/http://wiki.dbpedia.org/Datasets.

16. https://web.archive.org/web/20170331091345/http://folk.uio.no/bjornkho/MEP.

17. Ebenso wie bei der SMT basieren auch NMT-Ansätze auf statistischen Verfahren und nutzen damit ebenfalls die latent in den Sprachdaten enthaltenen Informationen. Aufgrund von Unterschieden in der internen Repräsentation der Daten ist eine terminologische Unterscheidung hier dennoch üblich.

18. Vgl. dazu z.B. die Modelle des OpenNMT-Systems (*Open Neural Machine Translation*, siehe https://web.archive.org/web/20181024202654/http://opennmt.net/Models/), einer *open-source*-Initiative der Harvard University in Zusammenarbeit mit Systran, einem der führenden Unternehmen im Bereich der Maschinellen Übersetzung.

Bibliographie

Van Aggelen, A. E.; Hollink, L.; Kemman, M.; Kleppe, M.; Beunders, H.. 2017. „The debates of the European Parliament as Linked Open Data." *Semantic Web – Interoperability, Usability, Applicability*, 8(2): 271–281. doi:10.3233/SW-160227.

Gale, W. and Church, K.. 1993. *A program for aligning sentences in bilingual corpora. Computational Linguistics*, 19(1).

Graën, J.; Batinic, D.; Volk, M.. 2014. „Cleaning the Europarl Corpus for Linguistic Applications". In: *Konvens 2014, Hildesheim, 8 October 2014 - 10 October 2014*. doi:10.5167/uzh-99005.

Kilgarriff, A.; Baisa, V.; Bušta, J.; Jakubíček, M.; Kovvář, V.; Michelfeit, J.; Rychlý, P.; Suchomel, V.. 2014. „The Sketch Engine: ten years on." *Lexicography*, 1: 7-36.

Koehn, P.. 2005. „Europarl: A Parallel Corpus for Statistical Machine Translation." *MT Summit* 5: 79-86.

Factsheet

Resource reviewed	
Title	Europarl
Editors	Philipp Koehn
URI	http://www.statmt.org/europarl/
Publication Date	2001-2012
Date of last access	19.01.2018

Reviewer	
Surname	Neuefeind
First Name	Claes
Organization	Institute for Digital Humanities, University of Cologne
Place	Cologne
Email	c.neuefeind (at) uni-koeln.de

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)	yes
Contributors	Are the contributors (editors, institutions, associates) of the project documented? (cf. Catalogue 1.3)	no
Contacts	Is contact information given? (cf. Catalogue 1.4)	yes
Aims		
Documentation	Is there a description of the aims and contents of the text collection? (cf. Catalogue 2.1)	yes
Purpose	What is the purpose of the text collection? (cf. Catalogue 2.2)	Research

Kind of research	What kind of research does the collection allow to conduct primarily? (cf. Catalogue 3.1.8)	Quantitative 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: Language technology
Content		
Era	What era(s) do the texts belong to? (cf. Catalogue 2.5)	Contemporary
Language	What languages are the texts in? (cf. Catalogue 2.5)	Danish, English, Finnish, French, German, Greek, Italian, Norwegian, Polish, Portuguese, Spanish, Swedish
Types of text	What kind of texts are in the collection? (cf. Catalogue 2.5)	Speech transcripts
Additional information	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)	yes
Selection	What selection criteria have been chosen for the text collection? (cf. Catalogue 3.1)	Language
Size		
Texts/records	How large is the text collection in number of texts/records? (cf. Catalogue 3.1.4)	> 1000
Tokens	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)	yes
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)	no
Text integration	What kind of material has been taken over from other sources? (cf. Catalogue 3.1.6)	none
Quality assurance	Has the quality of the data (transcriptions, metadata, annotations, etc.) been checked? (cf. Catalogue 3.1.7)	unknown
Typology	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)	Parallel corpus
Data Modelling		
Text treatment	How are the textual sources represented in the digital collection? (cf. Catalogue 3.2.1)	Orthographic transcription
Basic format	In which basic format are the texts encoded? (cf. Catalogue 3.2.4)	Plain text
Annotations		
Annotation type	With what information are the texts further enriched? (cf. Catalogue 3.2.2)	Structural information
Annotation integration	How are the annotations linked to the texts themselves? (cf. Catalogue 3.2.2)	Embedded
Metadata		
Metadata type	What kind of metadata are included in the text collection? (cf. Catalogue 3.2.3)	none
Metadata level	On which level are the metadata included? (cf. Catalogue 3.2.2)	not applicable
Data schemas and standards		
Schemas	What kind of data/metadata/annotation schemas are used for the text collection? (cf. Catalogue 3.2.4)	none

Standards	Which standards for text encoding, metadata and annotation are used in the text collection? (cf. Catalogue 3.2.4)	none
Provision		
Accessability of the basic data	Is the textual data accessible in a source format (e.g. XML, TXT)? (cf. Catalogue 4.1)	yes
Download	Can the entire raw data of the project be downloaded (as a whole)? (cf. Catalogue 4.2)	yes
Technical interfaces	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)	no
Reuse	Can you use the data with other tools useful for this kind of content? (cf. Catalogue 4.4)	yes
User Interface		
Interface provision	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)	no
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)	yes
Open Access	Is the text collection Open Access? (cf. Catalogue 6.2)	yes
Rights		
Declared	Are the rights to (re)use the content declared? (cf. Catalogue 6.2)	yes

License	Under what license are the contents released? (cf. Catalogue 6.2)	No explicit license / all rights reserved
Persistent identification and addressing	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)	no
Archiving of the data	Does the documentation include information about the long term sustainability of the basic data (archiving of the data)? (cf. Catalogue 6.4)	yes
Institutional curation	Does the project provide information about institutional support for the curation and sustainability of the project? (cf. Catalogue 6.4)	no
Completion	Is the text collection completed? (cf. Catalogue 6.4)	unknown
Personnel		
Editors	Philipp Koehn	