Lang Resources & Evaluation (2013) 47:1315–1326
DOI 10.1007/s10579-013-9231-6

PROJECT NOTE

Creating & Testing CLARIN Metadata Components

Folkert de Vriend · Daan Broeder · Griet Depoorter · Laura van Eerten · Dieter Van Uytvanck

Published online: 18 May 2013

Abstract The CLARIN Metadata Infrastructure (CMDI) that is being developed in Common Language Resources and Technology Infrastructure (CLARIN) is a computer-supported framework that combines a flexible component approach with the explicit declaration of semantics. The goal of the Dutch CLARIN project "Creating & Testing CLARIN Metadata Components" was to create metadata components and profiles for a wide variety of existing resources housed at two data centres according to the CMDI specifications. In doing so the principles of the framework were tested. The results of the project are of benefit to other CLARIN-projects that are expected to adhere to the CMDI framework and its accompanying tools.

Keywords Metadata · Infrastructure · CLARIN

F. de Vriend (✉)
Meertens Institute, Joan Muyskenweg 25, Amsterdam, The Netherlands
e-mail: folkert.de.vriend@meertens.knaw.nl

D. Broeder · D. Van Uytvanck
Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen, The Netherlands

D. Broeder
e-mail: Daan.Broeder@mpi.nl

D. Van Uytvanck
e-mail: Dieter.vanUytvanck@mpi.nl

G. Depoorter · L. van Eerten
Institute for Dutch Lexicology, Matthias de Vrieshof 2-3, Leiden, The Netherlands

G. Depoorter
e-mail: Griet.Depoorter@inl.nl

L. van Eerten
e-mail: Laura.vanEerten@inl.nl

1 Introduction

Descriptive metadata are used to characterize data resources and tools, to facilitate discovery and management in large (virtual) infrastructures and repositories. One of the goals of the Common Language Resources and Technology Infrastructure (CLARIN) project is to create a joint metadata domain for all Language Resources and Tools (LRT) (Váradi et al. 2008). To achieve this purpose a metadata infrastructure is being developed that combines a flexible component approach with the explicit declaration of semantics. This framework is called CLARIN Metadata Infrastructure (CMDI) and is described in Broeder et al. (2008). The need for a flexible component based metadata framework resulted from the experience in the LRT community that fixed schema solutions hamper broad usage due to the different needs and terminologies of subcommunities. This component metadata framework allows users, subcommunities and projects to design their own metadata schemas as long as they make use of widely agreed upon concepts that are stored in the ISOcat data category registry (Kemps-Snijders et al. 2009) and therefore guarantee interoperability.[1,2]

This paper will report on the Dutch CLARIN project "Creating & Testing CLARIN Metadata Components". The goal of this project was to create metadata components and profiles for a wide variety of existing resources housed at two data centres according to CMDI specifications. In doing so the principles of the metadata framework were tested. Since the results of the project became available in an early stage of CLARIN, they are of benefit to other CLARIN-projects that are expected to adhere to the CMDI framework and its accompanying tools. The project had three partners: The Max-Planck Institute for Psycholinguistics (MPI) carried out the coordination and management of the project.[3] The Institute for Dutch Lexicology (Instituut voor Nederlandse Lexicologie: INL) and the Meertens Institute (MI) were the two CLARIN-NL data centres that house the resources for which new CMDI metadata components and profiles were created and tested. Since INL and MI also aspired to become an official CLARIN data centre for which adherence to CMDI is a technical requirement, for these centres the project functioned as a preparatory phase as well.

In Sect. 2 we first summarize the basics of CMDI for creating and using metadata. In the rest of the paper the focus is on issues in using the CMDI principles for creating metadata for resources. In Sect. 3 we first describe what resources were selected at the data centres. In Sect. 4 we then discuss two dimensions across which a resource needs to be analyzed before metadata can be created. In Sect. 5 we go into detail about the actual creation of metadata components, profiles and records. Finally, in Sect. 6 we draw some conclusions.

2 The CMDI infrastructure

Although the principles of CMDI have been described in Broeder et al. (2008), we summarize the basics of CMDI and its terminology in this section.

[1] http://www.clarin.eu/files/metadata-CLARIN-ShortGuide.pdf.

[2] http://www.isocat.org.

[3] http://www.mpi.nl.

The CMDI design and construction was started by the European CLARIN (CLARIN EU) project to overcome the limitations of the existing metadata sets developed by the Open Language Archives Community (Simons and Bird 2008), the ISLE Meta Data Initiative (Broeder and Wittenburg 2006) and the Text Encoding Initiative (TEI).[4] Within the CLARIN EU project the MPI is responsible for guiding the implementation of CMDI and therefore is very interested in participating in projects that will use or test CMDI, like the project described in this paper.

CMDI uses ensembles of metadata components that are called profiles to create XSD metadata schemas that can be used to describe resources or collections of resources. Every metadata component is a set of metadata elements that is supposed to describe a specific aspect of a resource, e.g. an "actor" component specifies the biographical information of a person or a "location" component specifies the place where an event occurred. Every metadata element can link (using URI's) to a concept in a recognized data category registry such as the ISOcat data category registry or the Dublin Core Metadata Initiative (DCMI) metadata terms. Components and profiles are stored in the CMDI component registry so others can reuse them.

Instantiated schemas describe actual resources and are called metadata descriptions or metadata records. An important desideratum for CMDI is that it be flexible enough for any researcher to decide what metadata fit his or her needs best. The basics of CMDI and its terminology are depicted in Fig. 1.

In the CMDI infrastructure the metadata records are harvested with the OAI-PMH protocol and stored in a joint metadata repository. A CMDI service provider will then offer services like metadata search and browsing based on the repository's content to the world.

3 Data centres and resources

The MI data centre studies and documents Dutch language and culture.[5] Its main fields of research are ethnology and variationist linguistics. The research group Dutch Ethnology studies the dynamics and diversity in everyday cultural expressions. The research group Variationist Linguistics studies variation in the Dutch language as manifested in dialects, sociolects and ethnic Dutch. Most of the MI resources are accessible online, some together with computer tools for resource based research (see for instance Barbiers et al. 2007 or Meder 2010).

The INL data centre collects and studies Dutch words, stores them in databases—along with various additional linguistic data—and uses them to make scholarly dictionaries.[6] The INL also hosts the Dutch-Flemish Human Language Technology Agency (HLT Agency), which manages, maintains and distributes Dutch digital language resources for research, education and commercial purposes (Beeken and

[4] http://www.clarin.eu.

[5] http://www.meertens.knaw.nl.

[6] http://www.inl.nl.

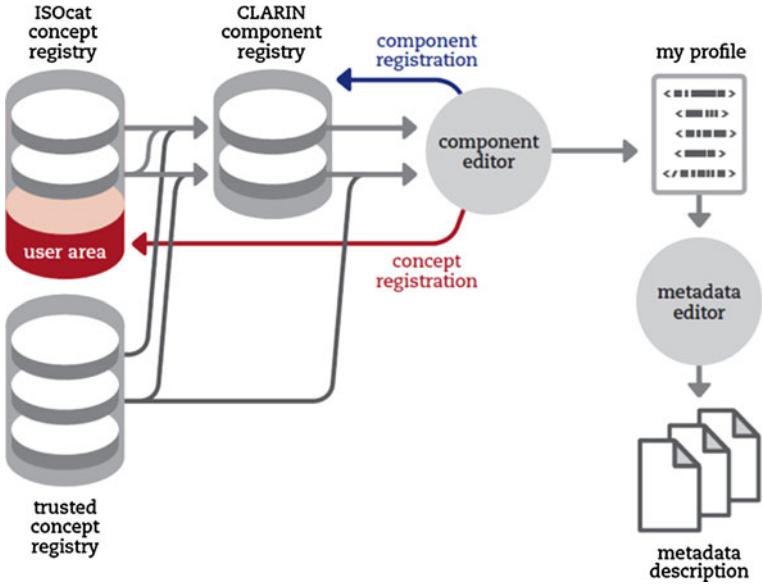

Fig. 1 Creating a CMDI metadata description

van der Kamp 2004).[7] Many of the resources available through the HLT Agency were developed by third parties.

In the project, metadata components were created for a sub selection of the many resources housed at the two data centres.[8] A strong preference was given to those resources at MI and INL that were non multi-media or multi-modal type of resources. These types of resources are, for instance, lexical resources or text corpora. For the multi-media and multi-modal type of resources it was expected that the existing component set and profile that was derived from IMDI would already be sufficient. After all IMDI was specifically aimed at describing this type of language resources.[9] Another reason to choose resources that are not typically described with IMDI metadata was that the main advocates of CMDI are from the same group that was instrumental in the development of IMDI.

The complete list of resources that were selected at the two data centres can be found in Tables 1 and 2 in the "Appendix" section. As can be seen the resources vary greatly. For MI we selected lexical resources (of proper names), linguistic databases (with syntactical, morphological and phonological dialect variation) and ethnological databases (containing data about folktales, songs, probate inventories and pilgrimages). For INL we selected lexical resources (monolingual and bilingual lexica, historical and scientific dictionaries), corpora (spoken and written) and

[7] The HLT Agency (http://www.inl.nl/tst-centrale), which is an initiative of and is funded by, the Dutch Language Union, was set up to organize easy access and re-usage of language resources for the Dutch language developed with public funding.

[8] Although CMDI can also be used for creating metadata for tools and web services, in the project these were not taken into account.

[9] http://www.mpi.nl/IMDI/.

historical documents (bible texts). The tables also indicate whether a resource has the characteristics of typical IMDI resources. This is indicated with either a 1 for "IMDI-like" or a 0 for "non-IMDI-like".

4 Resource analysis

The primary goal of CMDI is to enable the creation of adequate metadata components and profiles that have sufficient expressive power for the researcher to describe all relevant aspects of a resource. This can be a challenge when it is a new type of resource for which no metadata are available yet. In those cases a resource first needs to be properly analyzed.

4.1 Data versus metadata

Data other than the raw data of a resource (audio, video, etc.) could potentially be used as metadata. The following two types of such data are rather common:

- Very general data about the raw data that are used for data management purposes, for instance the ID of a recording.
- Data containing interpretations of the raw data, like a description of the transcription system used for recordings or a specialized (scientific) classification of recordings. An example of the latter would be a classification based on syntactic phenomena present in recordings ("agreement", "double negation", etc.).

We also need to distinguish between three main types of metadata, each one specific for a certain purpose (NISO 2004):

- Descriptive metadata describing a resource for purposes such as discovery and identification. Descriptive metadata can include elements such as "title", "abstract", "author", and "keywords".
- Structural metadata indicating how compound objects are put together, for example, how pages are ordered to form chapters.
- Administrative metadata providing information to help manage a resource, such as when and how it was created, its file type and who can access it.

CMDI metadata are primarily used as descriptive metadata, i.e. for discovery and identification of a resource (or parts of it). So when analyzing the aforementioned two types of data for their usefulness as metadata, this is the purpose that should guide one in deciding on what data to use as metadata. For instance, data about the location(s) of recordings in a resource could be valuable CMDI metadata for that resource, since it is plausible that a researcher searching for resources to use for his or her research is interested only in data that are bound to a specific geographic region. Data that are very specific for the corpus or database, such as for instance extensive "recording protocols" in speech corpora, are better stored as special information resources.

In principle no intrinsic distinction exists between the data and metadata of a resource. Some data can be used as metadata depending on the context that they function in, like resource discovery or data administration. One will have to look at

the context wherein the metadata will be used and be pragmatic with respect to the benefits and costs involved. Creating metadata long after the resource was created is always expensive.

A guiding principle in deciding what data should be described first is how useful the data are for researchers from other disciplines than the research discipline from which the resource originated. The CMDI infrastructure encourages reuse of resources by researchers from any sub discipline in the humanities or social sciences. Therefore, metadata that are useful to any researcher when browsing or searching for resources is especially valuable and should be focused on first. An example of such metadata would be "location of the recording" since it describes a very generic characteristic. The more specific characteristics, like for instance "syntactic attribute set used for adjectives", should be described after the generic characteristics have been fully documented.

Once a resource has been analyzed for data that can double as metadata one needs to see what other metadata are still needed. These metadata will have to be created from scratch.

4.2 Levels of granularity

An analysis of the levels of granularity present in the data of a resource (if any) is also needed. A resource can be a complex resource that can be (recursively) subdivided into constituents. Think of a text corpus that can be divided into subcorpora that can again be divided into individual texts. However, for resources in a relational database it is not always clear what the most suitable level of granularity is. For instance, when a resource consists of recordings for 100 different locations on 10 different subjects (10 syntactic phenomena for instance), what then is the most suitable granularity of these data? Are there 100 subcollections based on the 100 locations or are there 10 subcollections based on the 10 subjects? Neither is the "true" or "inherent" granularity for this resource. Again, in choosing one over the other one should be led by functional criteria. For example, what subcollections should be visible when searching for resources, or which ones should be citable? Or should one be able to transfer a certain subcollection to another repository?

For the MI resources, data granularity levels were assigned to the linguistic databases. For the INL resources, data granularity levels were assigned to most text and speech corpora. For the other MI and INL resources no functionally motivated data granularity levels could be assigned during the project. In Sect. 5 we will discus an example of a speech corpus (JASMIN) for which granularity levels were assigned.

When the levels of granularity for a resource have been decided upon metadata can be assigned to the constituents of the resource. When a description of a resource that can be subdivided into constituents (e.g. corpus, sub-corpus, text file) is distributed over several metadata records, decisions have to be made whether to duplicate metadata for the different metadata records. Although duplication takes care that the individual records are self-sufficient, it can lead to consistency problems. The CMDI infrastructure provides relations (links) between related metadata records e.g. between a collection and a sub collection, or between a text

corpus and a single text. These relations are specified by "isPartof" and "hasPart" links that are embedded in the metadata records. Duplicated metadata will then be redundant, since searching for metadata on e.g. text level also gives access to the metadata on corpus level.

5 Creating metadata components, profiles and records

After the selected resources (see the "Appendix" section) had been analysed, metadata components, profiles and records could be created. At the start of the project, the CMDI framework offered an initial set of ready-made metadata components, some of which were (partly) derived from existing metadata sets like OLAC, TEI, IMDI and DC. This initial set was created by the CMDI project for interoperability with the huge installed base of metadata records found in the LRT world. CMDI also offers a so-called XML-Toolkit to create CMDI metadata. Using this toolkit, users can create components using a standard XML editor in which schemas are used to enforce correctness and subsequently XSLT style sheets are used to create an actual CMDI metadata XSD schema. An XML editor is used to generate records for individual resources or sub collections.[10] This method of creating metadata was applied in the project reported on here. However, it is not very user friendly since it requires knowledge of XML, which most researchers do not have, and the procedure is cumbersome in itself. The CLARIN project since has also developed a set of user-friendly tools for creating components, profiles and records that greatly improve the usability of CMDI.[11]

Some metadata components created in the project could be derived from the set of components that were already available in the CMDI framework at the start of the project. These components contained either very general metadata elements (e.g. "location", "language") or metadata elements that were specifically intended to describe multi-modal (IMDI type) resources. Often the components were too limited or too detailed which then required creating a new component re-using some of the elements of the existing metadata components but incorporating others.

New components had to be created for the non-IMDI-like type resources. One example is the component "headwordtype" that is used for describing the headword type of a lexical resource at INL or MI ("lemma", "word form", "phrase" or "sentence"). Another example is the "dimensions" component that can be used for profiling a resource following the general research dimensions "time", "space" and "social". It describes for what research dimensions variation is present in the resource. For instance, it can describe that a resource contains social variation data for the social variables "religion", "age" and "gender".

In the components all newly introduced metadata elements were linked to data categories in the ISOcat data category registry (Kemps-Snijders et al. 2009). Each metadata element refers via a URI to exactly one data category in the data category registry (DCR), thus indicating unambiguously how the content of a metadata

[10] http://www.clarin.eu/toolkit.

[11] http://www.clarin.eu/cmdi.

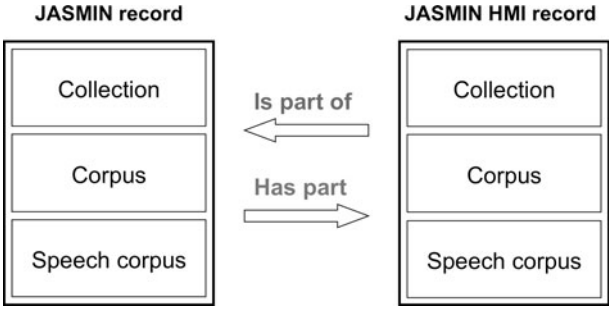

Fig. 2 Hierarchy and granularity in JASMIN metadata records

element should be interpreted. At the start of the project 217 data categories were available in the DCR. Mapping onto existing data categories, where possible, is strongly encouraged. Only if one is sure that the existing categories are not accurate enough should new ones be added to the DCR. Examples of data categories that were newly added during the project were "legal owner" and "pseudonym".

When trying to map the metadata elements onto existing data categories several issues arose. Sometimes the data concept definition given in the registry was too specific or too narrow. For the element "birth year" for example there was no concept "birth year" available in the DCR. There was however a concept called "birth date" which is related but not similar. We encountered the same issue for the concept "(overall) quality of the recordings of a speech corpus" for which only the related concept "quality of a recording" was available in the DCR. In these cases the decision was made to refer to the existing definitions, rather than to create new concept links. In other cases the definition of a term deviated too strongly from the definition that was envisioned when creating an element. For those concepts new data categories were created.

Another problem in relation to the DCR was double data categories (i.e. data categories with the same name and the same definition). In those cases the favoured data category was the one that was already in the ISOcat standardisation process.[12] If the data categories had the same status, preference was given to the data category that belonged to the Thematic Domain Group on metadata.

Newly created components can be combined in metadata profiles that can then be used to create metadata records for resources. Figure 2 illustrates the hierarchical structure of the metadata records that were created for the JASMIN resource (Cucchiarini et al. 2008). Each record first contains very general metadata at the collection level. These are then followed by more specific metadata at the corpus and speech corpus levels. All of the metadata records (and profiles) that were created in the project consisted of such hierarchal structures.

The different levels of data granularity of the JASMIN resource are also reflected in the metadata records depicted in Fig. 2. The whole JASMIN resource also contains a sub collection with speech for human machine interaction purposes called

[12] The standardisation process is carried out by domain experts who evaluate data categories and work in Thematic Domain Groups. Each Thematic Domain Group focusses on a specific topic like morphosyntax, metadata or lexicology.

JASMIN HMI. The arrows depict how a metadata record for the whole JASMIN corpus and a metadata record for the JASMIN HMI corpus refer to each other using "isPartof" and "hasPart" links.

6 Conclusions

The CMDI approach strongly encourages reuse of existing metadata components to avoid a proliferation of (similar) components in the component registry. However, when existing components are insufficient for proper metadata description of a resource it is also possible to create new components from scratch or to adapt existing components to one's specific needs. When creating new components it is encouraged to link them to the ISOcat data category registry. In the project reported on here CMDI appeared flexible enough for creating semantic descriptions of the resources at MI and INL. We were able to create components for both IMDI and non-IMDI-like resources using CMDI. What data granularity levels to discern when making existing resources available through the CMDI infrastructure should be functionally motivated.

All components developed in the project are available in the component registry.[13] The data categories created can be found in the ISOcat data category registry.[14] Finally, the project also published a Best Practice Guide for using CLARIN metadata components.[15]

Acknowledgments The authors would like to thank Jan Pieter Kunst (Meertens Institute) and Anna Aalstein (INL) for their valuable input during the project. The project reported on in this paper was funded by CLARIN-NL (www.clarin.nl).

Appendix

See Tables 1 and 2.

Table 1 Selected resources at MI

Diversity in Dutch DP Design (1)	A linguistic database with elicited speech and text collected between 2005 and 2009 to chart the syntactic variation at the level of nominal groups in the Netherlands, Belgium and North-West France
Dutch Database of Family Names (0)	A lexical resource containing an online dictionary and reference work for users interested in the origins, meanings and areas of distribution of surnames. The database contains 93,466 names
Dutch Database of First Names (0)	A lexical resource containing some 20,000 first names including explanations of the names and all sorts of other information on names

[13] http://catalog.clarin.eu/ds/ComponentRegistry/#.

[14] http://www.isocat.org/interface/index.html.

[15] http://trac.clarin.nl/raw-attachment/wiki/WikiStart/BestPracticeGuide-V4.pdf.

Table 1 continued

Dutch Songs Online (0)	An ethnological database with song texts from the Digital Library of Dutch Literature (DBNL) merged with metadata from the Dutch Song Database
Dynamic Syntactic Atlas of the Dutch dialects (1)	A linguistic database of elicited speech and text collected between 2000 and 2005 to chart the syntactic variation at the clausal level in 267 dialects of Dutch spoken in the Netherlands, Belgium and North-West France
Goeman Taeldeman Van Reenen project (1)	A linguistic database of elicited speech and text collected between 1980 and 1995 to chart morphological (word-level) variation
Pilgrimage in the Netherlands (0)	An ethnological database containing data about 662 pilgrimage centres. The data are relevant for research into pilgrimage, devotions to saints, religious material culture and religion in general
Plant names in Dutch dialects (0)	A lexical resource containing the popular names of plants in the Dutch language area. The database contains more than 275,000 records and is the world largest collection with this type of information
Probate Inventories Database (0)	An ethnological database containing 2,889 probate inventories from 10 places in the Netherlands dating from the seventeenth and eighteenth century
Soundbites (1)	A linguistic database containing more than 1,000 h of sound recordings of dialect speakers in over 100 places in the Netherlands. These recordings were collected in the 1950s, 1960s and 1970s
The Dutch Folktale Database (0)	An ethnological database that enables one to search for historical and contemporary fairy tales, legends, saints' lives, jokes, riddles and urban legends. Currently it contains 40,224 stories
The Dutch Song Database (0)	An ethnological database with over 125,000 songs from the Middle Ages to the modern times. The sources are songbooks, song sheets (broadsides), song manuscripts and fieldwork recordings

Table 2 Selected resources at INL

AUTONOMATA Spoken Names Corpus (1)	A speech corpus containing ca. 5,000 read-aloud first names, surnames, street names, city names and control words. The corpus consists of a Dutch part and a Flemish part
COREA Coreference Corpus (0)	A text corpus, which contains Dutch texts in which coreference relations were systematically marked. The corpus consists of newspaper, transcribed spoken language and lemmas from a medical encyclopedia
Corpus of Old Dutch (0)	A (historical) text corpus, which contains all the Dutch appellative word material that originated from the period 475–1200
Dictionary of Old Dutch (0)	A scientific and historical dictionary, which contains over 2,200 official documents from the thirteenth century
Dutch Electronic Lexicon of Multiword Expressions (0)	A monolingual lexical resource that contains more than 5,000 Dutch multiword expressions (MWEs). MWEs with the same syntactic pattern are grouped in the same equivalence class
Dutch PAROLE lexicon (0)	A monolingual lexical resource, which consists of about 20,200 entries, distributed over 13 parts of speech (POS). The entries have been described along the dimensions of morphosyntax and syntax

Table 2 continued

Eindhoven Corpus (0)	A text corpus containing the VU version of the Eindhoven Corpus (also: Corpus Uit den Boogaart). It is a collection of Dutch written and (transcribed) spoken texts from the period 1960–1976
e-Lex (0)	A monolingual lexical resource of Dutch consisting of a one-word lexicon and a multi-word lexicon. The one-word lexicon contains ca. 220,000 entries and more than 600,000 word forms, annotated with morphological, syntactical and phonological information
JASMIN speech corpus (1)	A speech corpus which contains ca. 115 h of Dutch speech by young speakers, non-native speakers and elderly speakers living in Flanders and the Netherlands
OMBI Dutch-Arabic (0)	A bilingual lexical resource, which contains ca. 35,000 entries. Dutch is the source language and Arabic the target language
PAROLE Corpus 2004 (0)	A text corpus, which contains modern Dutch texts (ca 20 million tokens), for the greater part originating from newspaper or magazine articles
Reference Lexicon for Belgian-Dutch (0)	A monolingual lexical resource of Belgian-Dutch containing ca. 4,000 entries (lemmas). The resource contains only those words, which have a specific meaning in Belgian-Dutch or appear only in Belgian-Dutch
Reference Lexicon for Dutch (0)	A corpus-based monolingual lexical resource of Dutch containing ca. 45,000 entries (lemmas). Apart from examples, each word has been provided with detailed linguistic information
Statenvertaling 1637 (0)	A historical document, which consists of the digitized texts of the bible Statenvertaling 1637
Woordenboek der Nederlandsche Taal (0)	A historical, scientific and descriptive dictionary of Dutch as used in 1500–1976
Wordlist of the Dutch Language 2005, source file (0)	A monolingual lexical resource with more than 100,000 entries annotated with linguistic information. The entries are high-frequency words of Standard Dutch and/or words that may cause spelling problems
5 Million Words Corpus 1994 (0)	A text corpus of ca. 5 million words derived from books, magazines, newspapers and TV broadcasts from the period 1970–1994
38 Million Words Corpus 1996 (0)	A text corpus, which consists of three main sub-corpora: a varied corpus (1970–1995), a newspaper corpus (Meppeler Courant, 1992–1995) and a legal corpus (1814–1989)

References

Barbiers, S., Cornips, L. & Kunst, J. P. (2007). The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas. In J. C. Beal & K. C. Corrigan & H. Moisl [red.] Creating and digitizing language corpora. Volume 1: Synchronic databases. Palgrave Macmillan, Hampshire, pp. 54–90.

Beeken, J. C. & van der Kamp, P. (2004). The Centre for Dutch Language and Speech Technology (TST Centre). In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pp. 555–558.

Broeder, D., Declerck, T., Hinrichs, E., Piperidis, S., Romary, L., Calzolari, N., & Wittenburg, P. (2008). Foundation of a component-based flexible registry for language resources and technology. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Broeder, D., & Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies, 1*(2), 119–132.

Cucchiarini, C., Driesen, J., Van Hamme, H., & Sanders, E. (2008). Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: The JASMIN-CGN Corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

ISLE Metadata Initiative (IMDI). (2009). Metadata Elements for Catalogue Descriptions. Part 1 B, Version 3.0.13. http://www.mpi.nl/IMDI/documents/Proposals/IMDI_Catalogue_3.0.0.pdf.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P. & Wright, S.E. (2009). ISOcat: Remodeling Metadata for Language Resources. In the special issue on the Open Forum on Metadata Registries of the *International Journal of Metadata, Semantics and Ontologies* (IJMSO), *4*(4), pp. 261–276.

Meder, T. (2010). From a Dutch Folktale Database towards an International Folktale Database. In: Fabula 51, Heft 1/2. Walter de Gruyter: Berlin: New York.

NISO. (2004). Understanding Metadata. Bethesda, MD: NISO Press. URL: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.

Simons, G., & Bird, S. "OLAC Metadata". 2008, cited version http://www.language-archives.org/OLAC/metadata-20080531.html, latest version http://www.language-archives.org/OLAC/metadata.html.

TEI Text Encoding Initiative. (2009). http://www.tei-c.org/.

Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.