Metadata Collection Records for Language Resources

Henk van den Heuvel*°, Erwin Komen°, Nelleke Oostdijk*
*CLST / CLS, Radboud University
°Humanities Lab, Radboud University
Erasmusplein 1, Nijmegen, the Netherlands
{h.vandenheuvel, e.komen, n.oostdijk}@let.ru.nl

Abstract

In this paper we motivate the need for introducing more elaborate and consistent metadata records for collections of (linguistic) data resources in the CLARIN context. For this purpose we designed and implemented a CMDI profile. We validated the profile in a first pilot in which we populated the profile for 45 Dutch language resources. Given the complexity of the profile and special purpose requirements we developed our own interface for creating, editing, listing, copying and exporting descriptions of metadata collection records. The requirements for this interface and its implementation are described.

Keywords: LR infrastructure; metadata; collection records; Collection Bank

1. Introduction

In the context of CLARIN (Common Language Resources Infrastructure)[1] various stakeholders have been working towards the realisation of an integrated, interoperable research infrastructure. Apart from the technological achievements and the services offered, key factors to the success of this infrastructure are the availability and accessibility of the language resources. Ideally the infrastructure is populated with as many resources as possible. Therefore continued efforts are being put into gathering what resources there are and curating them, targeting also the many resources that have been created in the context of research projects but have not (yet) found their way to one of the data centres or repositories.[2]

In order to provide an entry-point to the language resources available in the CLARIN infrastructure, the Virtual Language Observatory (VLO) was developed (Van Uytvank 2014; Van Uytvank et al. 2010). The VLO offers a faceted browser which allows users to search for resources through their metadata. User experiences with the VLO have shown that discovering resources, and especially resources of which one is not aware they exist, is problematic. From the analysis of Odijk (2014) we know that while some of the problems arise from the current limitations of the VLO, in many cases the cause of the problem lies in the nature of the metadata. The set of facets used for search in the VLO is (too) small, not all facets are relevant for the discovery of resources, and some facets are lacking and should be added.[3] In 2015 the CLARIAH Metadata Curation Taskforce[4] was charged with the task to come up with a profile for collection records that would support the search for and discovery of language (data) resources. Collections in this context are to be understood as creator or depositor defined aggregations of data that fulfil a certain purpose, They have been created explicitly so as to form a browsable and therefore manageable hierarchy (cf. Broeder et al. 2009: 51). An example of a collection is the Spoken Dutch Corpus (Oostdijk, 2000). It comprises some 800 hours of sound recordings along with various types of transcriptions and annotations, a lexicon and a number of frequency lists.

In the present paper we describe how we proceeded and arrived at a profile for collection records which we think overcomes most of the shortcomings that various profiles that have previously been used exhibited. Here we restrict ourselves to the metadata of (linguistic) data collections, thus excluding software and tools.

The structure of the remainder of this paper is as follows: In Section 2 we present the profile we developed for collection records. We motivate its design and the considerations underlying it. Then, in Section 3 we report on the pilot that we conducted. In this pilot we applied the profile to a select but varied set of resources. Next, we introduce the Collection Bank, an interface that we developed for entering and maintaining metadata collection records. This paper concludes with a summary of the main outcomes, and suggestions for future work.

2. Profile for Collection Records

The development of the profile was guided by a number of desiderata that we derived from Odijk (2014). Odijk found that

(a) often metadata elements that were crucial for the discovery of a resource were lacking as they were not mandatory
(b) values for several important metadata elements are not restricted to a closed set
(c) the metadata created by various researchers and research groups often display what he calls ‘unnecessary differences’
(d) the granularity of the metadata records varies wildly and is often too small.

[1] https://www.clarin.eu/

[2] In the Netherlands the Data Curation Service was set up as a centre of expertise to assist researchers in preparing their data for delivery to one of the CLARIN centres (van den Heuvel et al. 2015).

[3] Lušicky and Wissik (2017) also note the need for additional metadata. while using the VLO for discovering resources in the field of translation studies and urge other user groups to assess the VLO from their perspective and specify what should be added to satisfy their needs.

[4] In the Dutch CLARIAH project (https://www.clariah.nl/en/), the Metadata Curation Taskforce is concerned with providing metadata for various resources.

From this we concluded that we needed to establish what metadata elements are essential for search and discovery of resources, that such elements should be mandatory and to the extent possible, should have values from a closed/controlled vocabulary.

We started out by making an inventory of CLARIN-NL and CLARIN-EU collections represented in the CLAPOP[5], VLO[6], and EASY[7], and the information that was available in the form of collection records as well as any other information that might be considered relevant. It appeared that the information varied widely as can be seen from the comparison between CLAPOP, VLO and EASY in Table 1.

	CLAPOP	VLO	EASY
Title[8]	+	+	+
Research domain[9]	+		+
Annotations	+		
Format	+	+	
Resource tags	+		
Language	+	+	
Clarin centre[10]	+	+	
Country	+	+	
Resource type		+	
Availability		+	
National project		+	
Modality		+	
Organization		+	
Keyword		+	
Creator			+
Description			+
Subject			+
Coverage			+
Identifier			+

Table 1: Metadata available for resource discovery

We also looked at what profiles occurred in the CLARIN Component Registry. Here again there was a great deal of variation as most profiles appear to be collection-specific by design.

In developing our profile for collection records we opted to use the Dublin Core Metadata Element Set as a kernel similarly to what Bird and Simons (2003) did for OLAC. This decision was motivated by the fact that Dublin Core (DC)[11] is well-established: over the years it has been widely adopted and has proven to be usable for describing a wide range of resources. More specifically, it provides explicit definitions for each of the elements it contains. This, we expect, will contribute to the standardization we aim for. An illustration of where we deviate from DC is the distinction we make between the *type* and *subtype* of a resource. Thus in line with DC, the element *type* is defined as the nature of a resource, and follows the attributes defined in DC.[12] The attribute values associated with it, viz. ‘collection’, ‘dataset’, ‘image’, ‘sound’ and ‘text’ form a closed set. However, where in DC for example ‘dataset’ refers to any data encoded in a defined structure, our *subtype* for ‘dataset’ opens up a new closed set of attributes which distinguishes between ‘list’[13], ‘table’, ‘lexicon’ and ‘treebank’.

Since we were working in the CLARIN context, we found it opportune to select and copy the building blocks for our metadata collection profile from CLARIN’s component metadata.[14] Important profiles from which blocks were copied (and pruned and modified where needed) were: OralHistoryInterviews, SpeechCorpus, and textCorpusProfile.

From the start we were aware that the collection records were to be included in an interface where they could be searched. In determining the relevance of various metadata elements we made a clear distinction between those metadata elements that are relevant as search or filter criteria and those that are only informative. The elements we deemed most relevant for search and filtering purposes are: *title, type, modality, annotation type, temporal and geographical provenance*, and *language*.

Some of the metadata have a range of fixed values (closed sets). The permitted values are given in CLARIN’s Concept Registry[15] but, in CLARIN context, these are not (or rather: no longer) considered restrictive. Consequently, we added extra values for some of the metadata elements. This we did, for example, for *language* (where we needed a distinction for Northern and Southern Dutch (=Flemish)), and also for *annotation format*, *annotation type*, and *genre*.

Whenever available, the metadata elements in the profile have links to CLARIN’s Concept Registry.

Relations between collections are formulated via a specific provision since the relation category is not well supported in the Concept Registry and would lead to somewhat forced CMDI constructs. Moreover, many of the attributes available for existing types of relationship in DC and DataCite[16] do not cover the relations that we would like to describe such as *isSiblingOf*, *inRepository*, *hasSubset*. These are quite typical relations between resources, but these attributes are not available in DataCite and DC. These relations have therefore been provided in a separate text file where they are formulated as RDF tuples. This file is referred to in the CMDI metadata file.

The relations are embedded in the resource proxy list of the CMDI file and are illustrated in Figure 1 for one of the collections (the CGN 1.0). Furthermore we added metadata elements for other resource proxies such as search pages (URLs providing a search interface for a resource) and landing pages (URLs providing basic information about a resource). These are also shown in Figure 1.

[5] http://dev.clarin.nl/clarin-data-list-fs

[6] https://vlo.clarin.eu/

[7] https://easy.dans.knaw.nl/ui/?wicket:interface=:3::::

[8] In CLAPOP it is not possible to filter on title. VLO here uses ‘collection’.

[9] In EASY ‘audience’ is used.

[10] In VLO ‘data provider’ is used.

[11] http://dublincore.org/

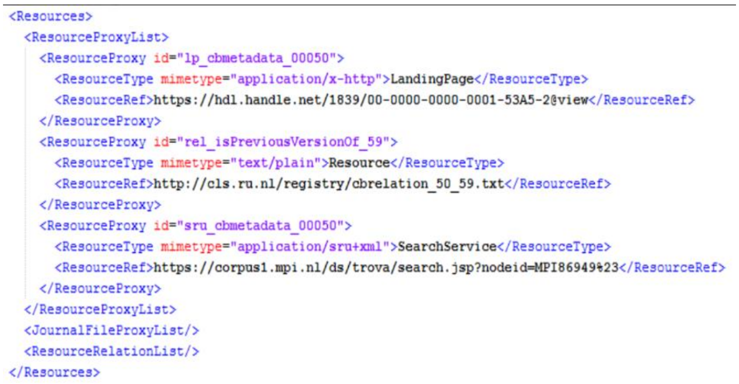
[12] http://dublincore.org/documents/dces/

[13] With ‘list’ also the type of list (frequency or other) and type of list items (word type, lemma, POS tag, n-gram, phrase type, sentence type, other) are associated as attributes.

[14] See https://www.clarin.eu/content/component-metadata and https://catalog.clarin.eu/ds/ComponentRegistry

[15] https://www.clarin.eu/ccr

[16] https://www.datacite.org/



```
<Resources>
  <ResourceProxyList>
    <ResourceProxy id="lp_cbmetadata_00050">
      <ResourceType mimetype="application/x-http">LandingPage</ResourceType>
      <ResourceRef>https://hdl.handle.net/1839/00-0000-0000-0001-53A5-2@view</ResourceRef>
    </ResourceProxy>
    <ResourceProxy id="rel_isPreviousVersionOf_59">
      <ResourceType mimetype="text/plain">Resource</ResourceType>
      <ResourceRef>http://cls.ru.nl/registry/cbrelation_50_59.txt</ResourceRef>
    </ResourceProxy>
    <ResourceProxy id="sru_cbmetadata_00050">
      <ResourceType mimetype="application/sru+xml">SearchService</ResourceType>
      <ResourceRef>https://corpus1.mpi.nl/ds/trova/search.jsp?nodeid=MPI86949%23</ResourceRef>
    </ResourceProxy>
  </ResourceProxyList>
  <JournalFileProxyList/>
  <ResourceRelationList/>
</Resources>
```

Figure 1: The relation component embedded in the resource proxy list

The text file expressing the relations then looks like this:

```
Collection    Type of relation     Collection
CGN1.0        isPreviousVersionOf  CGN2.0
```

A complete profile for our metadata collection record is stored in CLARIN's Component Registry and can be viewed at https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.1/profiles/clarin.eu:cr1:p_1493735943947/xml.

3. Pilot data for collection records

As a proof of concept for the sustainability of the profile that we conceived we conducted a pilot in which we used the profile to create the metadata collection records for 45 language resources. These were selected on the basis of the following criteria. The resource was (a) a Dutch language resource, (b) considered relevant for current linguistic research, (c) referenced in CLARIN-NL's CLAPOP[17] and (d) contained in the VLO[18] and/or LINDAT[19] but underspecified at collection level. Moreover, it was required that sufficient metadata information sources were available to make a (more or less) complete collection record. An overview of collections involved in the pilot is shown in the Appendix.

After selection of the resources, the metadata for the individual resources had to be retrieved and entered into the collection records for each resource. This work was carried out by a student assistant who was provided with project websites where metadata information per resource could be retrieved. Her work was supervised by one of the authors of this paper who also added URLs for search pages, landing pages, and the links to other (versions of the) databases to the record.

The student assistant started out with one Excel file per language resource in which she collected the metadata. Due to the hierarchy in the components of the metadata this approach soon faced its limitations. Moreover, the information should not only be stored in Excel format, but also be made accessible in CMDI metadata files. Therefore, we decided to look for an interface which allowed editing hierarchical metadata in a user friendly way whilst providing a CMDI file export option as well.

4. An Interface for Entering Metadata Collection Records

CLARIN offers a versatile and well documented metadata editor for CMDI profiles: COMEDI[20] (Lyse et al., 2014). This tool takes a CMDI profile file as input and allows entering values for each metadata record contained in it. However, the tool is very general in its set-up, whereas what we needed was a more specific metadata editor for our collection records, allowing us to include

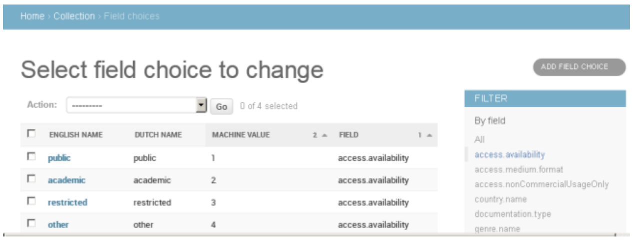
- clarifications per metadata category
- not yet existing components and metadata values
- a deeper hierarchical design of our components
- our own PIDs.

Moreover, edited metadata records should not be visible to everyone before being released. Based on these considerations we decided to develop an interface that would meet our requirements: the 'Collection Bank'.

4.1 The Collection Bank

The Collection Bank is a web application built on the Django framework.[21] The application facilitates creating, editing, listing, copying and exporting descriptions of metadata collection records.

The database model, which is hidden from the user, follows the 'CorpusCollection' specification of the model that is publically available in the CMDI Component Registry.[22,23] Saved collections can be 'published', which means that their *xml* representations become available through a persistent identifier. Saved collections become part of the user's list of collections (Collections > View).



Select field choice to change

Action: ---------- Go 0 of 4 selected

□	ENGLISH NAME	DUTCH NAME	MACHINE VALUE 2 ▲	FIELD 1 ▲
□	public	public	1	access.availability
□	academic	academic	2	access.availability
□	restricted	restricted	3	access.availability
□	other	other	4	access.availability

FILTER
By field
All
access.availability
access.medium.format
access.nonCommercialUsageOnly
country.name
documentation.type
genre.name

Figure 2: Administrator's interface to modify a fixed set

Changes in the definition of CorpusCollection in the CMDI Component Registry necessitate adapting the web interface. Changes in the fixed sets of metadata choices, for instance (e.g. *availability* can have the values 'academic', 'public', 'restricted' and 'other'), require the web interface's administrator to change the corresponding Field choices, as illustrated in Figure 2.

[17] http://portal.clarin.nl/clarin-data-list-fs

[18] https://vlo.clarin.eu/

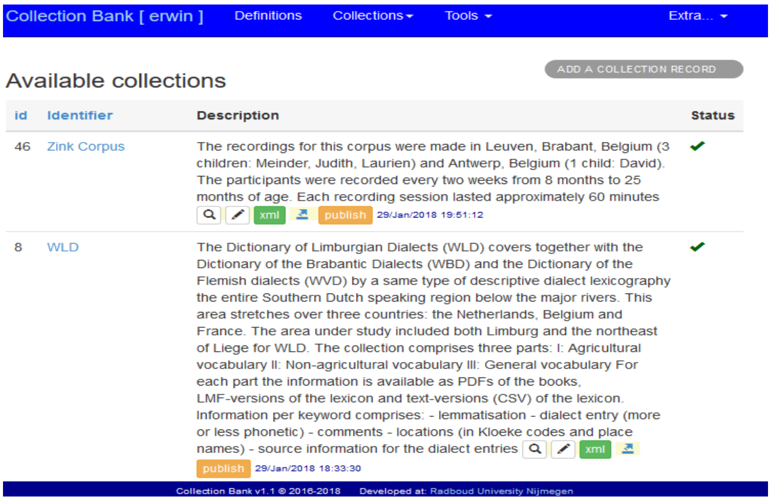
[19] https://lindat.mff.cuni.cz/repository/xmlui/discover

[20] http://clarino.uib.no/comedi/

[21] See http://applejack.science.ru.nl/collbank. An account can be created through Extra > SignUp. The program heavily uses the existing facilities within Python's Django package.

[22] https://catalog.clarin.eu/ds/ComponentRegistry/

[23] CorpusCollection: clarin.eu:cr1:p_1493735943947



Available collections

id	Identifier	Description	Status
46	Zink Corpus	The recordings for this corpus were made in Leuven, Brabant, Belgium (3 children: Meinder, Judith, Laurien) and Antwerp, Belgium (1 child: David). The participants were recorded every two weeks from 8 months to 25 months of age. Each recording session lasted approximately 60 minutes xml publish 29/Jan/2018 19:51:12	✓
8	WLD	The Dictionary of Limburgian Dialects (WLD) covers together with the Dictionary of the Brabantic Dialects (WBD) and the Dictionary of the Flemish dialects (WVD) by a same type of descriptive dialect lexicography the entire Southern Dutch speaking region below the major rivers. This area stretches over three countries: the Netherlands, Belgium and France. The area under study included both Limburg and the northeast of Liege for WLD. The collection comprises three parts: I: Agricultural vocabulary II: Non-agricultural vocabulary III: General vocabulary For each part the information is available as PDFs of the books, LMF-versions of the lexicon and text-versions (CSV) of the lexicon. Information per keyword comprises: - lemmatisation - dialect entry (more or less phonetic) - comments - locations (in Kloeke codes and place names) - source information for the dialect entries xml publish 29/Jan/2018 18:33:30	✓

Collection Bank v1.1 © 2016-2018 Developed at: Radboud University Nijmegen

Figure 3: List-view of collections

The menu option ‘list-view’ of the collections (illustrated in Figure 3) provides an overview of all collections that have been entered. Each collection comes with its publication status and date (if published) and offers the user a number of action buttons: view, edit, download, copy and publish. The *view* button opens a human readable view of the collection, the *edit* button opens the editor (see below). The *download* button allows exporting a collection as a CMDI *xml* file (the Tools > Export commands facilitate exporting all the user’s records together). The *copy* action allows copying whole metadata collection records (easing the work on similar collections). The last button *publish* creates the CMDI *xml* file, adding a persistent handle to it. All collections can be viewed and downloaded outside of the CollBank application through their persistent handle.[24]

Validation of the record's contents is done after creation of its CMDI *xml* representation; it is part of both the download as well as the publish actions. Validation is done through the *xsd* definition of a collection record, which is made available in the CLARIN components registry.

Upon creation of a record (Collections > Add), the collection *editor* opens up, and the user can specify the obligatory and optional parts of the metadata collection.[25]. Figure 4 (see next page) contains parts of the entry for the VU-DNC (Vis, 2011). The user needs to provide a named identifier for the collection (which is used for easy referencing within the web application). The 'description' field allows adding a description of any length to the collection as a whole. Note the clickable help-link below the text-input field ('See: Description…').

The VU-DNC collection contains two 'resources', which become visible on the same page by clicking the 'Show' buttons. The principle behind the Collbank web application is that the fields of one collection are all accessible on one web page. Figure 4 does not show the other editable collection fields, but it does contain the form's bottom row that holds the different save and delete options available to the collection entry as a whole.

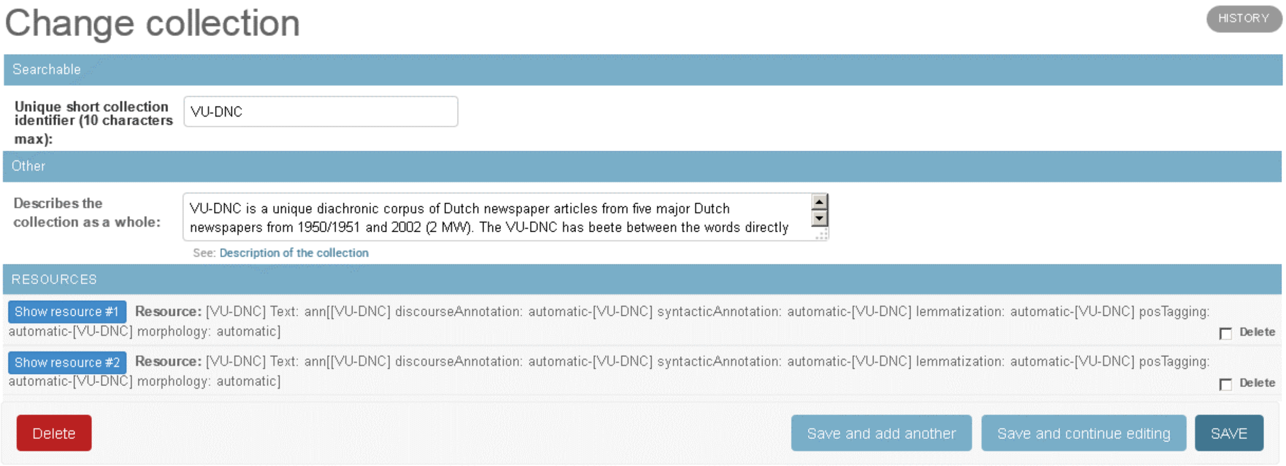
5. Conclusion and Future Work

In this contribution we have provided the motivation for introducing more elaborate and consistent metadata records for collections of (linguistic) data resources. We have presented the CMDI profile which we developed for this purpose and presented the selection of collections used for populating the profile. Given the complexity of the profile and special purpose requirements we developed our own interface for creating, editing, listing, copying and exporting descriptions of metadata collection records.

As a follow-up of the work presented here, the metadata collection records will be made available and searchable in CLARIN’s CLAPOP portal and in the VLO. The Collection Bank interface will be used to add further metadata collection records in the near future.

[24] Browsing to a persistant handle results in a text-oriented summary of the collection metadata, while the *xml* version of the record is returned in other situations. The specification of the VU-DNC collection, as an example, is available at http://cls.ru.nl/registry/cbmetadata_00016, or with its PID, at http://hdl.handle.net/21.11114/COLL-0000-000B-CAD5-1.

[25] Error handling, e.g. where obligatory fields are not filled in, follows the standards set by the Django framework.



Unique short collection identifier (10 characters max): VU-DNC

Describes the collection as a whole: VU-DNC is a unique diachronic corpus of Dutch newspaper articles from five major Dutch newspapers from 1950/1951 and 2002 (2 MW). The VU-DNC has beete between the words directly

See: Description of the collection

Resource: [VU-DNC] Text: ann[[VU-DNC] discourseAnnotation: automatic-[VU-DNC] syntacticAnnotation: automatic-[VU-DNC] lemmatization: automatic-[VU-DNC] posTagging: automatic-[VU-DNC] morphology: automatic] □ Delete

Resource: [VU-DNC] Text: ann[[VU-DNC] discourseAnnotation: automatic-[VU-DNC] syntacticAnnotation: automatic-[VU-DNC] lemmatization: automatic-[VU-DNC] posTagging: automatic-[VU-DNC] morphology: automatic] □ Delete

Figure 4: Specifying details of one collection

6. Acknowledgements

The work reported here was funded by the Dutch CLARIAH programmme[26] under project number CC-WP3-15-002. We are grateful to Hanna van den Heuvel for editing the metadata collection records in the project.

7. References

Bird, S. and Simons, G. (2003). Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. *Computers and the Humanities*, 37: 375-388.

Broeder, D., Gaiffe, B., Gravilidou, M. Lemnitzer, L. Van Uytvanck, D., Witt, A., and Wittenburg (2009). *Metadata Infrastructure for Language Resources and Technology*. 2009-02-02. Version 5. Deliverable D2.4 from EC FP7 project no. 212230. Retrieved from https://www.clarin.eu/sites/default/files/wg2-4-metadata-doc-v5.pdf

Heuvel, H. van den, Oostdijk, N., Sanders, E., and Lint, V. de (2015). Data curations by the Dutch Data Curation Service : Overview and future perspective. In *CLARIN 2014 Selected Papers; Linköping Electronic Conference Proceedings* # 116, pp. 54-62. http://www.ep.liu.se/ecp/116/005/ecp15116005.pdf

Lušicky, V. and Wissik, G. (2017). Discovering resources in the VLO: A pilot study with students of translation studies. *Selected papers from the CLARIN Annual Conference 2016. Linköping Electronic Proceedings* 136: 63-75.

Lyse, G.I., Meurer, P. and De Smet, K. (2014). COMEDI: A new component metadata editor. In *Proceedings of the CLARIN Annual Conference 2014*: https://www.clarin.eu/sites/default/files/cac2014_submission_13_0.pdf

Odijk, J. (2014). *Discovering resources in CLARIN: Problems and suggestions for solutions*. Working Paper. http://dspace.library.uu.nl/handle/1874/303788

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first evaluation. In *Proceedings of the second International Conference on Language Resources and Evaluation (LREC2000),* http://www.lrec-conf.org/proceedings/lrec2000/pdf/110.pdf

Oostdijk, N. and H. van den Heuvel (2014). The evolving infrastructure for language resources and the role for data scientists. In *Proceedings of the ninth International Conference on Language Resources and Evaluation (LREC2014)*, pp. 608-612. http://www.lrec-conf.org/proceedings/lrec2014/index.html

Van Uytvanck, D. (2014). How can I find resources using CLARIN? Presentation held at the Using CLARIN for Digital Research tutorial workshop at *the 2014 Digital Humanities Conference, Lausanne, Switzerland*. https://www.clarin.eu/sites/default/files/CLARIN-dvu-dh2014_VLO.pdf. July 2014

Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., and Gardellini, M. (2010). Virtual Language Observatory: The portal to the language resources and technology universe. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC2010),* pp. 900-903. http://www.lrec-conf.org/proceedings/lrec2010/

Vis, K. (2011). Subjectivity in News Discourse. A corpus linguistic analysis of informalization. *PhD thesis*, VU Amsterdam. https://research.vu.nl/ws/portalfiles/portal/2925809

8. References to Language Resource Portals

[CLAPOP] re3data.org: CLAPOP; editing status 2017-01-30; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R35884

[Collection Bank] http://applejack.science.ru.nl/collbank

[LRT inventory] http://www.clarin.eu/view_resources and http://www.clarin.eu/view_tools

[VU-DNC] https://portal.clarin.inl.nl/vu-dnc/

[26] https://www.clariah.nl/en/

APPENDIX: Collections included in the pilot set

Name collection	Relevant information available from
Corpus Gesproken Nederlands (CGN)	http://lands.let.ru.nl/cgn/ https://corpus1.mpi.nl/ds/asv/?0&openpath=node:2102153 http://tst-centrale.org/nl/tst-materialen/corpora/corpus-gesproken-nederlands-detail
SoNaR	https://dev.clarin.nl/node/4195 http://link.springer.com/book/10.1007%2F978-3-642-30910-6 http://tst-centrale.org/nl/tst-materialen/corpora/sonar-corpus-detail
D-LUCEA	https://portal.clarin.nl/node/4183 https://corpus1.mpi.nl/ds/asv/?0&openpath=node:2102153# http://lucea.wp.hum.uu.nl/summary/ http://dev.clarin.nl/clarin-data-list-fs
LESLLA	https://corpus1.mpi.nl/ds/asv/?0&openpath=node:2102153# ! http://dev.clarin.nl/clarin-data-list-fs
Dictionary of the Brabantic Dialects, Dictionary of the Limburgian Dialects, Dictionaries of the dialects of Gelderland	http://www.lrec-conf.org/proceedings/lrec2016/pdf/223_Paper.pdf http://dialect.ruhosting.nl/wbd/index.htm http://dialect.ruhosting.nl/wld/index.htm http://dialect.ruhosting.nl/wgd/index.htm
MIMORE Data: DiDDD MIMORE Data: DynaSAND MIMORE Data: GTRP	http://portal.clarin.nl/node/4213
VALID	http://validdata.org/ especially: http://validdata.org/clarin-project/datasets/ https://corpus1.mpi.nl/ds/asv/?0&openpath=node:2102153 Papers: DOI:10.1075/dujal.3.2.02heu http://www.lrec-conf.org/proceedings/lrec2014/index.html at Authors van den Heuvel
VU-DNC	http://portal.clarin.nl/node/4194 https://portal.clarin.inl.nl/vu-dnc/
Academia Collectie (NIBG)	http://portal.clarin.nl/node/4230 https://www.academia.nl/ -> https://www.academia.nl/faq/28341 https://vlo.clarin.eu/search?0&fq=collection:Nederlands+Instituut+voor+Beeld+en+Geluid+Academia+collectie
DBD/TCULT	http://www.clarin.nl/sites/default/files/IDCC13-DCS_v4.2-final.pdf https://corpus1.mpi.nl/ds/asv/?openpath=node:84720 https://www.clarin.eu/sites/default/files/cac2014_submission_15_0.pdf
DiscAn	https://dev.clarin.nl/node/4198 https://tla.mpi.nl/resources/discan-corpora/ http://dx.doi.org/10.1016/j.dcm.2012.09.003 https://corpus1.mpi.nl/ds/asv/;jsessionid=9C64195BC53CEFED79D1B544E8822C23?0
DUELME Data	http://portal.clarin.nl/node/4200 http://duelme.clarin.inl.nl/ http://duelme.clarin.inl.nl/documentation.php https://vlo.clarin.eu/record?2&docId=hdl_58_10032_47_270633b99d34b5fc06b0699e8e2dd93c&fq=collection:TST-Centrale&index=1&count=2 Go to Advanced: Show all metadata fields
INTERVIEWS Data	https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:41923 https://dev.clarin.nl/node/4201
IPNV	https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:46232 https://www.clarin.eu/sites/default/files/cac2014_submission_15_0.pdf

LAISEANG	https://dev.clarin.nl/node/4197 https://corpus1.mpi.nl/ds/asv/;jsessionid=9C64195BC53CEFED79D1B544E8822C23?0
NEHOL	https://dev.clarin.nl/node/4193 https://corpus1.mpi.nl/ds/asv/;jsessionid=9C64195BC53CEFED79D1B544E8822C23?0 https://hdl.handle.net/1839/00-0000-0000-0016-83D9-D@view
ETC Database	https://portal.clarin.nl/node/4180 https://shebanq.ancient-data.org/sources https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:48490/tab/1 https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:58245/tab/1
Liederenbank	http://www.liederenbank.nl/index.php?lan=en https://nl.wikipedia.org/wiki/Nederlandse_Liederenbank
IFA Corpus = IFA speech corpus = IFA Spoken Language Corpus	https://vlo.clarin.eu/record;jsessionid=2E26AC7EC25FC5DDDA76EF19B781A537?1&docId=hdl_58_1839_47_00-0000-0000-0003-46DA-E&q=IFA+corpus&fq=languageCode:code:nld&index=0&count=5 http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFAcorpus/ https://corpus1.mpi.nl/ds/asv/?0
IFA Dialogue Video Corpus	https://vlo.clarin.eu/record?1&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-576_64_format_61_cmdi&q=ifadvcorpus&index=0&count=2 http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/
Corpus NGT (Nederlandse GebarenTaal)	http://www.ru.nl/corpusngtuk/ http://www.ru.nl/corpusngt/de_filmpjes/download-filmpje/ (licenses) https://corpus1.mpi.nl/ds/asv/?4&openpath=node:319374
CHILDES Dutch corpora	http://childes.psy.cmu.edu/ https://vlo.clarin.eu/record?2&docId=http_58__47__47_hdl.handle.net_47_11372_47_LRT-439_64_format_61_cmdi&q=childes&fq=languageCode:code:nld&index=0&count=3692 http://childes.talkbank.org/access/Dutch/
ESF Corpus	https://user.clarin.eu/resources/mpi-esf-corpus https://corpus1.mpi.nl/ds/asv/?4&openpath=node:319374