Related Works in the Linguistic Data Consortium Catalog

Daniel Jaquette, Christopher Cieri, Denise DiPersio
Linguistic Data Consortium
3600 Market Street, Philadelphia, PA. 19104 USA
{jaquette, ccieri, dipersio}@ldc.upenn.edu

Abstract

Defining relations between language resources provides an archive with the ability to better serve its users. This paper covers the development and implementation of a Related Works addition to the Linguistic Data Consortium's (LDC) catalog. The authors go step-by-step through the development of the Related Works schema, implementation of the software and database changes, and data entry of the relations. The Related Work schema involved developing of a set of controlled terms for relations based on previous work and other schema. Software and database changes consisted of both front and back end interface additions, along with modification and additions to the LDC Catalog database tables. Data entry consisted of two parts: seed data from previous work and 2019 language resources, and ongoing legacy population. Previous work in this area is discussed as well as overview information about the LDC Catalog. A list of the full LDC Related Works terms is included with brief explanations.

Keywords: relations, metadata, language resources

1. Introduction

Categorizing relations between language resources offers the ability for an archive to better serve its users by identifying additional resources of potential interest, deepening a resource's metadata and increasing findability. There are multiple ways to implement this functionality; these include established metadata standards in META-SHARE, ISOCat, OLAC (Open Language Archives Community) and Dublin Core. The Linguistic Data Consortium (LDC) recently deployed such a metadata field in the LDC Catalog called "Related Works." Related Works are defined by a controlled vocabulary of relations based on the standards mentioned above with modifications relevant to the Consortium's language resources.

We describe the Related Works schema, and the steps to implementation. The latter include additional database functionalities to store and display Related Works information and the processes to recognize and add relations between/among the Consortium's language resources and external resources (where appropriate). As of this writing, Related Works have been cataloged for roughly seventy percent of LDC's holdings and the data indicates that each corpus has an average of about 2.5 relations. Even though not yet complete, the Related Works designation has enriched the Catalog by improving database infrastructure and providing users with additional, useful information about LDC's language resources.

2. The LDC Catalog

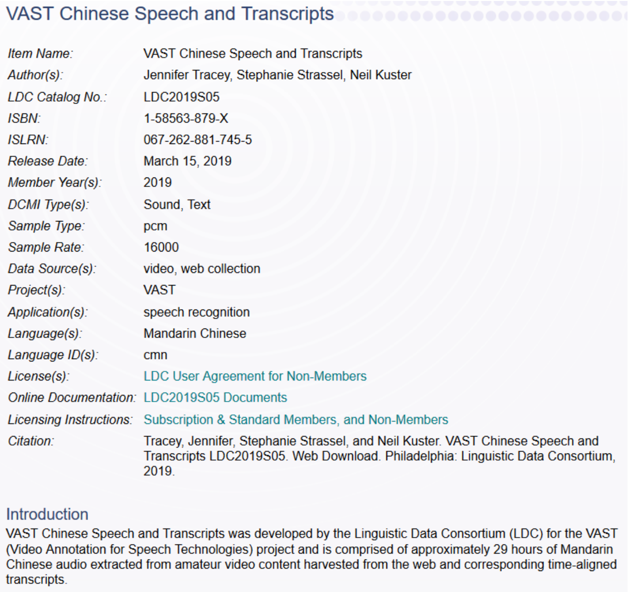
The LDC Catalog consists of over 800 publicly accessible Language Resources (LR), adding approximately 36 new ones each year. In the LDC Catalog, these LRs are data sets used for a variety of different applications for research, technology development, and instruction in the disciplines of human language technologies and linguistics. See Figure *1* for an example of part of a catalog entry.

3. Taxonomy of Relations among Language Resources

Labropoulou, Cieri and Gavrilidou (2014) proposed a taxonomy of relations among language resources to be incorporated into corpus metadata and documentation. The taxonomy was based upon a review of schema previously developed for META-SHARE and the ISOCat Data Category Registry and then tried against the LDC Catalog. Specifically, the authors reviewed the entries for each of the 574 corpora then included in the LDC Catalog. 337 of those made informal mention of relations to other datasets. An important discovery from this effort was: "*If we take this as representative of the field, it means that more than half of all data sets are related to one or more other data sets. This fact alone should make it clear why the study of LR relations is important to the field.*"

Each mention was then encoded in the proposed taxonomy. In many cases, the review led to the discovery of relations that were not mentioned in the Catalog in any way. In some cases, the facts of a specific pair of datasets led to changes in the taxonomy.

The paper identified future work including the application of META-SHARE/LDC taxonomy to the complete LDC Catalog and the effort to identify some relations automatically. The sections that follow describe the full implementation of the former and a small step toward the latter.



VAST Chinese Speech and Transcripts

Item Name:	VAST Chinese Speech and Transcripts
Author(s):	Jennifer Tracey, Stephanie Strassel, Neil Kuster
LDC Catalog No.:	LDC2019S05
ISBN:	1-58563-879-X
ISLRN:	067-262-881-745-5
Release Date:	March 15, 2019
Member Year(s):	2019
DCMI Type(s):	Sound, Text
Sample Type:	pcm
Sample Rate:	16000
Data Source(s):	video, web collection
Project(s):	VAST
Application(s):	speech recognition
Language(s):	Mandarin Chinese
Language ID(s):	cmn
License(s):	LDC User Agreement for Non-Members
Online Documentation:	LDC2019S05 Documents
Licensing Instructions:	Subscription & Standard Members, and Non-Members
Citation:	Tracey, Jennifer, Stephanie Strassel, and Neil Kuster. VAST Chinese Speech and Transcripts LDC2019S05. Web Download. Philadelphia: Linguistic Data Consortium, 2019.

Introduction

VAST Chinese Speech and Transcripts was developed by the Linguistic Data Consortium (LDC) for the VAST (Video Annotation for Speech Technologies) project and is comprised of approximately 29 hours of Mandarin Chinese audio extracted from amateur video content harvested from the web and corresponding time-aligned transcripts.

Figure 1: Part of a catalog entry for VAST Chinese Speech and Transcripts (Tracey et al. 2019).

4. Deviations from META-SHARE/LDC

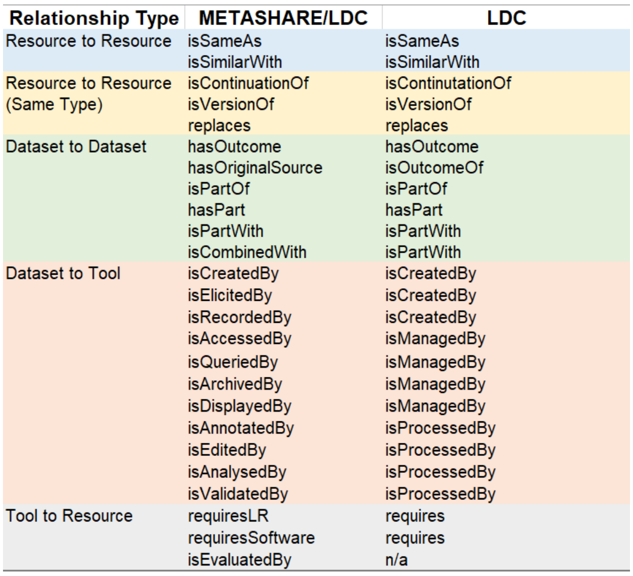
The META-SHARE/LDC schema discussed earlier provided a solid starting point for the development of our own schema. We deviated in a few ways from that proposed set of vocabulary to choose relations that best fit our current model of corpora relations. Specifically, they were as follows:

1. Where there was no proposed inverse relation, we added one. Additionally, for LDC's purposes it was appropriate to merge *hasOutcome* and *hasOriginalSource* into one symmetrical set of *hasOutcome/isOutcomeOf* terms. The granularity difference between the two original terms was not needed for the LDC Catalog.
2. The LDC catalog currently stores a very limited number of tools as records. For that reason, we consolidated the various "Dataset to Tool" relations into three terms: *isCreatedBy*, *isProcessedBy*, and *isManagedBy*. Similarly, we found of the "Tool to Dataset" relations, only *isRequiredBy* fit our purposes. In the event we decide to expand our tool-based catalog entries, and need the increased granularity, this decision can be revisited and terms updated accordingly.
3. Once we started applying the schema to the data, we found that an overwhelming number of corpora could be related with *isOutcomeOf* and so added a more specific refinement option with *isAnnotationOf*.
4. While *isSimilarWith* is a fairly broad relation, we required a term to cover two resources when no specific relation could be applied. Thus, we added the general term, *relatesTo*.

See Table *1* for a table mapping the proposed META-SHARE/LDC relations to LDC's Related Works.

Below is a list of unused META-SHARE/LDC relations. While there is some nuance lost in consolidation of terms, we felt the benefit of a smaller set of terms outweighed the benefit of specificity when applied to LDC's resources.

- Merged with isOutcomeOf
    - hasOriginalSource
- Merged with isPartWith
    - isCombinedWith
- Merged with isCreatedBy
    - isElicitedBy
    - isRecordedBy
- Merged into isManagedBy
    - isAccessedBy
    - isQueriedBy
    - isArchivedBy
    - isDisplayedBy
- Merged into isProcessedBy
    - isAnnotatedBy
    - isEditedBy
    - isAnalysedBy
    - isValidatedBy
- Merged into requires
    - requiresLR
    - requiresSoftware



Relationship Type	METASHARE/LDC	LDC
Resource to Resource	isSameAs	isSameAs
	isSimilarWith	isSimilarWith
Resource to Resource (Same Type)	isContinuationOf	isContinuationOf
	isVersionOf	isVersionOf
	replaces	replaces
Dataset to Dataset	hasOutcome	hasOutcome
	hasOriginalSource	isOutcomeOf
	isPartOf	isPartOf
	hasPart	hasPart
	isPartWith	isPartWith
	isCombinedWith	isPartWith
Dataset to Tool	isCreatedBy	isCreatedBy
	isElicitedBy	isCreatedBy
	isRecordedBy	isCreatedBy
	isAccessedBy	isManagedBy
	isQueriedBy	isManagedBy
	isArchivedBy	isManagedBy
	isDisplayedBy	isManagedBy
	isAnnotatedBy	isProcessedBy
	isEditedBy	isProcessedBy
	isAnalysedBy	isProcessedBy
	isValidatedBy	isProcessedBy
Tool to Resource	requiresLR	requires
	requiresSoftware	requires
	isEvaluatedBy	n/a

Table 1: Crosswalk between METASHARE/LDC terms and LDC terms.

5. Schema Development

As noted above, the first step in developing LDC's schema was to evaluate the proposed set of terms from the META-SHARE/LDC paper. Once we had established which terms worked best for our purposes, we developed a schema with definitions, usage instructions, and examples in a similar style to the Open Language Archives Community (OLAC) Metadata Set (Simons, 2008). Below is a list of schema relations with brief descriptions. Inverse relations are included on the same line for brevity's sake. The full schema contains more detailed usage instructions.

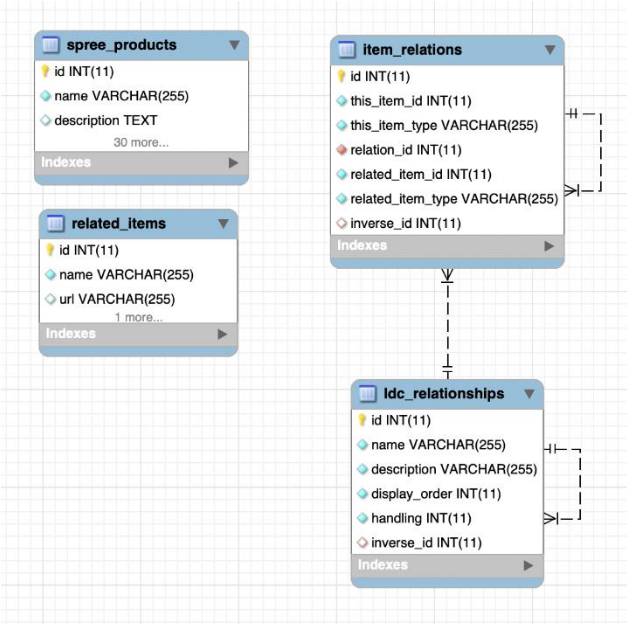
5.1 List of Relation Terms

- isSameAs
    - Resource A is the new/alternate name for A, while the content is identical.
- isSimilarWith
    - Resource A is similar to B in regards to creation specifications, purpose, source material, etc… or is part of a series.
- relatesTo
    - Resource A relates to B in some broad general manner.
- isContinuationOf / hasContinuation
    - Resource A continues the work of resource B.
- isVersionOf / hasVersion
    - Resource A is an extension in size, corrections of content, etc… of resource B.
- replaces / isReplacedBy
    - Resource A replaces or supersedes resource B.
- isOutcomeOf / hasOutcome
    - Resource A is the product/outcome of resource B.
- isAnnotationOf / hasAnnotation
    - Resource A is annotation of resource B.
- isPartOf / hasPart

    - Resource A is part of resource B.
- isPartWith
    - Resource A and B are both parts of a third resource.
- isCreatedBy / creates
    - Resource A was created by tool B.
- isProcessedBy / processes
    - Resource A was processed by tool B.
- isManagedBy / manages
    - Resource A is managed by tool B.
- requires / isRequiredBy
    - Tool A requires resource B.

6. Platform Implementation[1]

Once the schema was established, the next step was to implement the Related Works field in the catalog infrastructure. This involved two principal tasks: adding tables to the underlying relational database, and revising the front end display for catalog entries. Three tables were added to the database as illustrated in Figure *3*. While there was some potential benefit in using a graph database, the current business system in which the catalog resides already uses relational databases. For this reason, we sought to implement the needed databases as relations, and used a software solution to create the second half of a transitional database.



spree_products
id INT(11)
name VARCHAR(255)
description TEXT
30 more...
Indexes

related_items
id INT(11)
name VARCHAR(255)
url VARCHAR(255)
1 more...
Indexes

item_relations
id INT(11)
this_item_id INT(11)
this_item_type VARCHAR(255)
relation_id INT(11)
related_item_id INT(11)
related_item_type VARCHAR(255)
inverse_id INT(11)
Indexes

ldc_relationships
id INT(11)
name VARCHAR(255)
description VARCHAR(255)
display_order INT(11)
handling INT(11)
inverse_id INT(11)
Indexes

Figure 3: Related Works tables in MySQL

Relations essentially store three pieces of data: the resource, the type of relation, and the related resource. However, in order to support references to related works outside of the LDC Catalog, it was also necessary to store names and URLs for these exogenous resources. (Labropoulou, Cieri, and Gavrilidou, 2014). One example would be a corpus developed with a tool that lives outside the LDC Catalog.

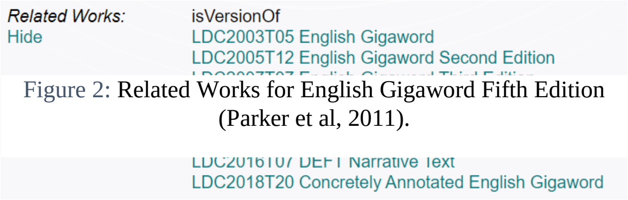
One exception in the storing of the relations in our database is that of *isPartWith*, which connects two resources that are related to a third resource via *isPartOf*. Rather than store each *isPartWith* relation in the same manner as the other relations, this relation is built on the fly on each catalog entry by querying for other resources that have the same *isPartOf* relation. We chose this implementation as the number of *isPartWith* entries grow exponentially with each new part. Due to this implementation, it was necessary, in some cases, to implement so-called "dummy resources" for cases where two resources are connected with *isPartWith* but the parent part does not exist as a whole. One use of this is for corpora that are distributed in parts due to size constraints; the whole corpus exists in theory, but there is no actual corpus or URL. For example, GALE Phase 3 Arabic Broadcast News Speech Part 1 (Walker, 2016) and GALE Phase 3 Arabic Broadcast News Speech Part 2 (Walker, 2017) should be related to each other with *isPartWith*, and so it is necessary under this approach system to create a hidden virtual resource (that one might think of as "GALE Phase 3 Arabic Broadcast News Speech Complete") in the database and relate each part to it with *isPartOf*.

The implementation also automatically fills in converse relations. For instance, if Corpus A *isAnnotationOf* Corpus B, then the system will automatically populate the database with an entry for Corpus B *hasAnnotation* Corpus A.

For the front-end display, we decided that related works should be viewable as a list on the Catalog in a way that satisfies user expectations and maximizes the user experience. The solution was to display the Related Works from most to least specific relation, with inter-corpora relations taking precedence over corpus-to-tool relations. Multiple relations of the same type are further sorted chronologically. See Figure *2* for an example Related Works section in the LDC Catalog.

7. Seeding Data Entry

To get started with the task of data entry for our 800 plus corpora, we used a spreadsheet that had been developed for use in writing the aforementioned META-SHARE/LDC paper. At the time of that writing, 337 of the 574 corpora made mention of another corpus in their description. Using the relations that had already been defined for that set, along with manual entry of relations for all 2019 corpora as a result, approximately 120 relations were used to seed the database, enabling a soft launch of Related Works in July of 2019.



Related Works: isVersionOf
Hide
LDC2003T05 English Gigaword
LDC2005T12 English Gigaword Second Edition
LDC2007T07 English Gigaword Third Edition
LDC2016T07 DEFT Narrative Text
LDC2018T20 Concretely Annotated English Gigaword

Figure 2: Related Works for English Gigaword Fifth Edition (Parker et al, 2011).

[1] LDC's Catalog/Business System is built using Spree, an e-commerce platform that runs on Rails.

8. Ongoing Data Entry

For resources dating back to 1993, two LDC staff members who did not participate in the development of the schema are evaluating each corpus manually for its relations. In many cases, the related corpus is easy to identify as LDC has made it a practice to link related works in the corpus description. Thus, for the majority of relations, the task is to find the hyperlinks to other resources and determine the relation. However, the corpus documentation is also reviewed to determine relations that may not be reflected in the catalog entry’s description.

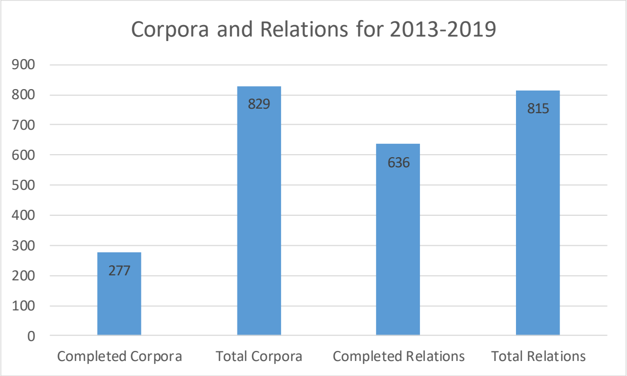


Figure 4: Chart showing total corpora and relations in LDC's catalog from 2013-2019. Each corpus averages 2.3 relations

Relations are submitted in batches for the LDC Publications group to review and edit. This process allows us to continue to refine the schema and thus increase the benefit to the user. The relations are then imported into the catalog and the process begins again. At the time of this writing, relations have been added for all corpora released since 2004, for a total of 585 corpora and 1,476 relations, with an average of approximately 2.5 relations per corpus. Total relations across all years currently amount to 1,545 relations and 840 corpora. See Figure *4* for these totals and Figure *5* for a breakdown of percentages of specific term use. We expect legacy data entry to be completed by the time this paper is presented at LREC2020.

In the course of the data entry project we found a few sets of corpora whose interconnectedness was more appropriately analyzed as a whole and these cases led to modifications to the schema. For example, LDC catalogs a number of corpora used for ongoing evaluations. At first glance, these may have been considered versions of each other. However, we felt given the data itself often changed dramatically with only the objective of the data set remaining the same, the continuation set of terms was better suited. This case, as well as a few others were then explicitly added to the schema to ensure consistent cataloguing moving forward.

9. Conclusion and Future Work

We have described the development of the Related Works metadata field and its implementation into the LDC Catalog. This new functionality makes it clear when a corpus is part of a series, or split into parts, or has a depth of derived works which could be of use or interest to a researcher. Additionally, the ability to automatically complete both sides of a relation eases the burden on the cataloger to capture needed information and reduces the probability of error.

LDC has not yet formally announced the addition of related works to the catalog metadata since work is ongoing as of this writing. Nevertheless, we asked a question about it in our 2020 member survey sent in February. A little under half of respondents were not yet aware of the new functionality, but of those that were, all but one reported finding it useful, with over a third reporting related works as “very useful.”

In terms of next steps, LDC is exploring whether the related works infrastructure can be extended to include papers written about LDC data, of which more than 10,000 have been identified thus far. (Ahtaridis, et al. 2012). In addition to demonstrating the research impact of LDC data, providing users with the ability to view, at a glance, the papers written about a particular corpus facilitates their understanding about the resource and may spark additional research ideas, enriching the community all the more.

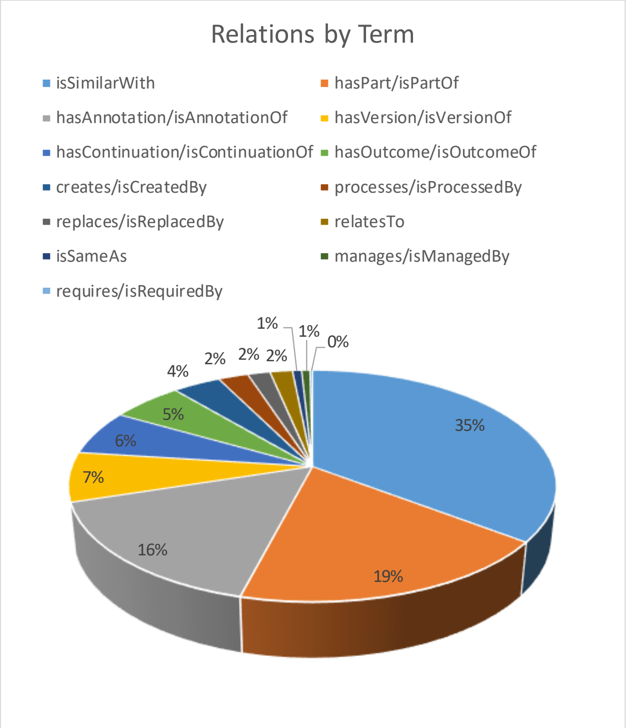


Figure 5: Chart showing relations by term, sorted most to least frequently occurring. Data is from 2004-2020. Note *isPartWith* is not represented in this chart since it is automatically generated.

10. Acknowledgements

The authors would like to thank Beth Nicholson for the software and database development of the Related Works module and Meghan Glenn and John Vogel for the legacy data entry aspect of the project.

11. Bibliographical References

Ahtaridis, E., Cieri, C., DiPersio, D. (2012). LDC Language Resource Papers: Building a Bibliographic Database. Eighth Edition of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, May. European Language Resource Association (ELRA).

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerq, T., Francopoulo, G., Arranz, V., Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. Proceedings of the Eighth International Conference on Language Resources (pp. 1090 - 1097). ELRA

Gavrilidou, M., Labropoulou, P., Piperidis, S., Speranza, M., Monachini, M., Arranz, V., Francopoulo, G. (2011). META-NET Deliverable D7.2.1 - Specification of Metadata-Based Descriptions for Language Resources and Technologies.

ISO 12620 (2009). Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources, International Organization for Standardization.

Labropoulou, P., Cieri, C., and Gavrilidou, M. (2014). Developing a Framework for Describing Relations among Language Resources. Ninth Edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, May. European Language Resource Association (ELRA).

Mariani, J., Cieri, C., Francopoulo, G., Paroubek, P., Delaborde, M. (2014). Facing the Identification Problem in Language-Related Scientific Data Analysis. Ninth Edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, May. European Language Resource Association (ELRA).

Simons, G., Bird, S. (2008). OLAC Metadata. Open Language Archives Community, Nov. 7, 2019, www.language-archives.org/OLAC/metadata.html

12. Language Resource References

Parker, Robert, et al. (2011). English Gigaword Fifth Edition, distributed via Linguistic Data Consortium, 1.0, ISLRN 911-942-430-413-0.

Tracey, Jennifer, et al. (2019). VAST Chinese Speech and Transcripts, distributed via Linguistic Data Consortium, 1.0, ISLRN 067-262-881-745-5.

Walker, Kevin, et al. (2016). GALE Phase 3 Arabic Broadcast News Speech Part 1, distributed via Linguistic Data Consortium, 1.0, ISLRN 597-417-124-701-7

Walker, Kevin, et al. (2017). GALE Phase 3 Arabic Broadcast News Speech Part 2, distributed via Linguistic Data Consortium, 1.0, ISLRN 141-827-463-794-4