Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 45 (2017) 375–391

www.elsevier.com/locate/csl

Supporting accessibility and reproducibility in language research in the Alveo virtual laboratory

Steve Cassidy*[,a], Dominique Estival[b]

[a] *Department of Computing, Macquarie University, Australia*
[b] *MARCS, Western Sydney Univeristy, Australia*

Received 2 September 2016; Accepted 13 January 2017

Abstract

Reproducibility is an important part of scientific research and studies published in speech and language research usually make some attempt at ensuring that the work reported could be reproduced by other researchers. This paper looks at the current practice in the field relating to the citation and availability of both data and software methods. It is common to use widely available shared datasets in this field which helps to ensure that studies can be reproduced; however a brief survey of recent papers shows a wide range of styles of citation of data only some of which clearly identify the exact data used in the study. Similarly, practices in describing and sharing software artefacts vary considerably from detailed descriptions of algorithms to linked repositories. The Alveo Virtual Laboratory is a web based platform to support research based on collections of text, speech and video. Alveo provides a central repository for language data and provides a set of services for discovery and analysis of data. We argue that some of the features of the Alveo platform may make it easier for researchers to share their data more precisely and cite the exact software tools used to develop published results. Alveo makes use of ideas developed in other areas of science and we discuss these and how they can be applied to speech and language research.

Keywords: Corpus infrastructure; Data citation; Reproducibility; Research methods; eResearch; Research workflow

1. Introduction

This paper is about reproducibility of research in the fields that make use of digital archives of written and spoken language data. The goal of the paper is to survey the current practice in making published research reproducible and then examine ways in which this might be improved. In particular, we claim that some of the features of the Alveo Virtual Laboratory are useful in making research studies easier to reproduce and extend.

Reproducibility has been an issue in science for many years with many studies looking at the benefits of replication (Muma, 1993) and the degree to which research results might be reproduced (Kelly et al., 1979). Recently, a survey paper in psychology showed that only one third to one half of the results of 100 studies could be fully reproduced (Open Science Collaboration, 2015); a result that has prompted much debate in this and other fields.

* Corresponding author.
E-mail address: steve.cassidy@mq.edu.au (S. Cassidy), d.estival@westernsydney.edu.au (D. Estival).

http://dx.doi.org/10.1016/j.csl.2017.01.003
0885-2308/

The research disciplines we discuss include acoustic phonetics, speech technology, natural language processing, lexicography, socio-linguistics and linguistics more generally, but also include aspects of psychology and musicology. These disciplines all make use of digital recordings of human language as audio and video recordings, textual transcripts and annotations and associated signals such as those capturing physiology and movement. In the core disciplines, few papers are written without some reference to a digital collection whether small or large. Collecting this data is often the most time-consuming part of any research project and it is to the credit of the field that this has led to the availability of many widely shared data-sets that are used as the raw material for research. While many data sets are shared and available through services like LDC[1] and ELRA[2], others are only available via download from a research group website and still others are held by researchers who may share data following a request or not depending on a multitude of factors.

Even once the *data* relating to a study is available, reproduction of the study relies on being able to reproduce the various data filtering, processing and analysis steps described in papers published about the study. While these may be clearly described, they are usually carried out using software tools and our ability to reproduce the results presented in the paper may depend on being able to run the same tools with the same input parameters. The researcher trying to reproduce a study then faces the problem of finding the right software, then the right version of the software, then a machine to run the software on that is compatible and finally finding the right settings to run the equivalent experiments.

A further consideration is the difficulty in getting some software tools to work, either when trying to reproduce a published result or just replicate a procedure described in a paper. Further to this, given the wide range of technical expertise of researchers in this field, there is an issue of accessibility for tools that might be impossible to run without deep technical knowledge. This may stand in the way of the reproduction of some work but also makes the application of tools in a cross-disciplinary manner difficult.

These are not new problems and they are not confined to one discipline; solutions have been developed steadily over the years to make data and tool sharing easier. The infrastructure that we now have available to us with large amounts of storage and fast network infrastructure opens up new possibilities that hold great promise in improving access to data and tools.

This paper describes the support provided by the Alveo system for sharing data and building reproducible workflows for speech and language research. We first provide a review of current practice with respect to data sharing and reproducibility in the field and in science more generally. We then review the range of data repositories that are available for language data. The paper then provides an overview of the Alveo platform with a commentary on how it seeks to address some of the problems raised in our reviews. We conclude with a discussion of how Alveo might contribute to improved practices in speech and language research.

2. Reproducibility in language research

This section provides a review of the current best practice in language research relating to the sharing and publication of data and tools and the reproducibility of research studies.

2.1. Access to data

Data is fundamental to speech and language research; the availability of large scale storage means that studying large quantities of speech data is now feasible whereas some years ago it would have been difficult.

The scale of data used in studies varies from recordings of a few speakers in a study of the physiology of speech production to the many thousands of speakers and hours of data used to train speech recognition systems.

To illustrate the treatment of data in our field we carried out a brief review of two volumes of this journal (Computer Speech and Language, Volumes 39–40, 2016) looking in particular at the sources of data used. There are eleven (11) papers in total in these two volumes covering a range of topics from generating speech from articulatory measurements to modelling Chinese word form patterns. All papers make reference to one or more collections of language data used in the study. Six (6) of the papers made use of text corpora while five (5) were based on speech

[1] https://www.ldc.upenn.edu/
[2] http://www.elra.info/en/

and articulatory data. Seven (7) of the papers used previously published data sets, four (4) used data harvested from Wikipedia and only one referred to a new data set created as part of the study although one other refers to an extended version of a previous CNN transcript corpus.

Four of the studies in these volumes make use of data harvested from Wikipedia (Calvo et al., 2016; Zhang et al., 2016; Qin et al., 2016; Zamani et al., 2016). Wikipedia is a common source of data for NLP research because it is a large collection with broad authorship covering a wide range of topics; in addition, it is available in multiple languages with some known relations between the different language editions. The issue with Wikipedia from a reproducibility perspective is that it is a constantly changing resource and so even snapshots taken days apart can vary significantly. To address this, there are a number of published snapshots of Wikipedia that have been made available specifically as NLP corpora, for example Reese et al. (2010) describes a pre-processed dump of three language versions of Wikipedia. Of the four papers using Wikipedia in this survey, all use different language versions (English, Swahili, Chinese and Persian) and it seems that all created the corpus as part of the study. In two cases, a download link is mentioned for the corpus; however only one of these links works at the time of writing (six months after publication of the studies). Only one paper gives a time that the snapshot was taken (November 2013) however it is not clear that even given this information it would be possible to reconstruct the dataset.

Two speech corpora are used in (Bayer and Riccardi, 2016) (WSJ and LUNA), the authors cite papers that describe these corpora rather than using catalogue identifiers. LUNA is available via ELRA and has a catalog reference; similarly, WSJ is available from LDC who provide a recommended citation that includes the catalogue identifier. Ghaemmaghami et al. (2016) makes use of the NIST SRE 2004 and 2008 data sets but gives no concrete citation of these. Similarly Lopez-Moreno et al. (2016) uses the NIST LRE'09 speech data set, citing the NIST evaluation plan but not the LDC catalogue identifier. These are well-known data sets in the research community and most researchers would be able to identify them from the name, however the LDC does publish a catalogue identifier for the datasets and again includes a suggested citations. These examples show the assumptions that are made about the availability and identifiability of well-known datasets. In all cases here, the specificity of the resource names is such that a researcher would be very likely to be able to obtain the same resource.

Prasad and Ghosh (2016) makes use of the MRI-Timit corpus which combines audio recordings and MRI scans. The corpus is cited with a reference to the paper describing the corpus (published in 2011) but does not include a reference to the data itself. However, the original paper does include a URL which still resolves to a working page containing information about acquiring the data.

Gonzalez et al. (2016) describes a method of synthesising speech from articulatory data captured using permanent magnet articulatography (PMA). Part of this work is the collection of a suitable set of data using a novel articulatography hardware setup, hence there is no existing data available that could be used in the study. This paper typifies another point in the data landscape where the data collection is an integral part of the innovation described in the study. The paper clearly describes the way that the data was collected but does not offer a downloadable version of the data for future researchers. This is not surprising or unusual; data like this is difficult to collect and the authors are understandably keen to make as much of it as possible before opening it up to other researchers. This model is similar to some of the other examples described here (for example the MRI-TIMIT corpus), with the only difference being that this is the first publication about this dataset. It is reasonably common for datasets like this to be made more widely available after a suitable quarantine period to allow the authors to exploit the data themselves.

This brief review of current practice shows the diversity of resources that are used in language and speech research. What is clear is that data is central to this research and every paper takes some care in describing the data collected or providing unambiguous names and citations that would allow another researcher to find the same data. While there are established means to refer to widely used data held by ELRA and LDC, the common practice is to cite a paper that describes the dataset or describes early work using the data; failing this, well-known names of the resources are used that can be resolved using a search engine to identify the data source and how to acquire it. In some cases where the data is newly collected for the study, it is made available for download, even if the download link is not working shortly after publication. We can conclude that researchers understand that the availability of data, or the ability to reconstruct a similar dataset, is an important part of a research publication in this field.

The weakness identified in this review comes from the imprecise way in which some researchers are referring to their data and the lack of detail in some cases on what filtering of data was done to select the exact inputs for the experiments. The general references to Wikipedia as a source allow us to understand what kind of data is being used and even to collect our own similar dataset, but does not mean that we have access to exactly the data used in the

study. In some cases, corpora are referenced but it is clear that some selection has been done on the data to remove outliers or noisy data such that the exact input for the experiment is not clear (even if the selection mechanism is easy to understand).

A final point to make is about the fragility of data that is shared by uploading it to a personal or institutional website. It is very easy to share data in this way, just by making an archive of the required files and making it available on an existing server; however, most websites are subject to periodic review, especially now that the web is often seen as a vehicle for marketing in Universities rather than for scholarly communication. These reviews rarely respect the "Cool URIs don't change" mantra (Berners-Lee, 1998) and so the previously published dataset becomes lost to future researchers. The lesson to learn here is that we need to make it easy for researchers to publish data in places that are committed to preserving access into the future. These may be institutional or disciplinary data repositories, but the goal must be that making data available is as easy for the researcher as it would be to place an archive file on a local web server.

2.2. Reproducible workflows

Scientific workflows are increasingly carried on in the digital realm and this is certainly true of workflows in the domain of language research. The research workflow is the series of steps that transforms the raw data (audio, video, text) into the results included in a published paper. This can involve automated steps such as formant tracking of speech signals or POS tagging of text and manual tasks such as annotation or tagging of entities in the data.

Workflow steps are carried out using some kind of software tool, whether this is fully automated or manual. The first step towards being able to reproduce a workflow is to have access to the software that was used in the study or an equivalent tool. However, even this is not the end of the story.

Cohen et al. (2016) describes the *hierarchy of needs* for reproducibility of a software based study, particularly in the domain of natural language processing:

1. Availability: the system must be available, or there must be sufficient detail available to reconstruct the system, exactly.
2. Builds: the code must build.
3. Runs: the built code must run.
4. Evaluation: it must be possible to run on the same data and measure the output using the same implementation of the same scoring metric.

This is formulated in the context of NLP studies where it is common for a paper to describe the results obtained by a single system that has been built to perform a given task such as *Entity Linking* or *Sentiment Analysis*. These systems often involve custom code although they may build on widely available components such as the Stanford NLP Tools (Manning et al., 2014). Cohen's hierarchy, which is similar to a number of others described in the more general computer science literature (Collberg et al., 2016), concentrates on the availability and workability of the software system and would be applicable to most of the papers in the CSL sample discussed earlier. However, in many cases the workflow for a language based research study makes use of a number of software components and might combine manual and automatic steps towards an end result.

To illustrate this alternate case we can look at an example study published in the Journal of Phonetics: *Classification of gender based on cepstral coefficients and spectral moments* (Spinu and Lilley, 2010). This paper follows a common workflow in this field that might be summarised as:

1. Select stimuli from an existing corpus.
2. Play recorded stimuli to listeners to elicit a categorisation.
3. Measure stimuli using acoustic features.
4. Apply classification methods over signals.
5. Generate plots and discuss.

Software tools are used in steps 3–5 in this workflow and understanding these steps fully would be a key requirement in being able to reproduce the study. In the paper, these steps are described largely in terms of the methods that

were used (*"first 6 cepstral coefficients Bark-scaled"*, *"we used HMMs to divide the segments into regions of internally minimized variance"*, etc.) rather than specific software and settings. In some cases, specific software is mentioned (Praat version 5.4.01, for spectrograms and spectra) but these are general packages rather than specific routines and settings.

This practice has the advantage that reproducing the workflow should be possible using any number of software packages capable of performing the same well-known computations. Most of the methods are well understood, deterministic computations that should be consistently reproduced by any software providing that parameterisation. In some cases however, there is more complexity – for example the use of HMMs to segment the acoustic signal. In this case, the particular package used and the details of the parameters used would be important in trying to reproduce the work. This is particularly important where the reproduction attempt generates different results to the original research; access to the same software and parameter settings can help clarify whether bugs in the software or procedural errors are the source of the discrepancies.

Based on these observations, we might suggest an extended hierarchy of needs for reproducible research in speech and language:

1. All computations clearly described in terms of well-known methods.
2. In addition, specific software packages are referenced.
3. and details of which functions within these were used.
4. and the exact settings used to run the computation.
5. and a copy of the scripts used for data processing is available.
6. and it can be downloaded and executed.
7. and the results are the same as those in the paper.

The assertion here is that the further down this list one can go for a particular study, the stronger the claim that the research described is reproducible. In some cases, the first requirement in the list will be sufficient to fully describe the research and most reviewers would be unhappy if a paper did not meet this standard.

2.3. Reproducibility in other disciplines

Reproducibility standards and practices vary across different disciplines in science and technology. It is informative to look at some areas where more effort has been made in recent years to ensure that published studies are reproducible and see if there is anything we can learn from their approaches.

As mentioned earlier, a recent paper in Science on the reproducibility of studies in Psychology Open Science Collaboration (2015) has triggered a lot of discussion in that discipline. As part of the follow-on from that study there is a move towards more transparent publication of research data and methods. The Transparency and Openness Promotion (TOP) Guidelines[3] developed by the Center for Open Science, attempt to set out increasing levels of best practice for journals with respect to open data and reproducible research. These include statements about data and code sharing, pre-registration of studies and the publication of replications of studies. A large number (714) of journals have signed up in support of these guidelines which may result in changes in the way that research is published in future.

The Center for Open Science, the sponsor of the Psychology Reproducibility Project, also provides tools for collaboration and data sharing that aim to support reproducibility as part of the scientific workflow. The Open Science Framework[4] (OSF) provides a shared workspace where collaborators can store data files and analysis code. The workspace supports version control for uploaded files and can be made public in whole or in part following the publication of a study. OSF encourages researchers to *fork* a project repository in order to reproduce the work; a forked repository contains all of the contents of the original and can then be used as the workspace for the reproduction effort.

[3] https://cos.io/top/
[4] https://osf.io/

Repositories for research data are becoming more common with services such as Figshare[5], Dryad[6] and Zenodo[7] providing storage space that can be referenced with a Digital Object Identifier (DOI) for citation in a journal publication. These repositories are intended to support sharing of intermediate results derived during data analysis for publication, hence they store spreadsheets, measurements and plots generated as part of the research workflow. This has particular value in addition to sharing the source data in a study, as it illustrates how the final analysis was arrived at and provides the opportunity for further levels of validation of the analysis that was carried out.

Another OSF initiative are a set of badges[8] to acknowledge open practices. Badges are awarded by journals against papers that fulfil certain criteria with respect to reproducibility. For example, the Open Data badge can be awarded for a paper that makes data relevant to the study available in a publicly accessible repository under an open access licence with a data dictionary sufficient to allow a third party to understand the data. In a recent survey, Kidwell et al. (2016) found that the use of these badges in the journal Psychological Science had resulted in a significant increase in the number of papers making data available compared with similar journals that did not use badges.

2.4. Reproducibility of workflows

In addition to sharing research data, there is a lot of work on making it easier to share reproducible research workflows. Since the heart of any computational research workflow is software, one focus of attention is to improve the practice of researchers in the way that they write software to support their work. The Software Carpentry initiative (Wilson, 2006) provides training to scientists in the use of scripting languages (Python, R) and version control to keep track of the scripts they write. Services like Github and Bitbucket are often used to host source code repositories but since they are not intended as long-term archival repositories, they are not suited for citation in themselves. Github recently launched an initiative in collaboration with Zenodo to allow a DOI to be assigned to a particular version of software from a Github repository[9].

While writing scripts to automate research analysis is possible for some researchers, it may not be an option for a large number who don't have the technical background required to set up the required software and author the code themselves. To support this class of user, a number of graphical workflow engines have been built that allow an interactive workflow construction. The most widespread of these are Taverna (Wolstencroft et al., 2013) and Galaxy (Afgan et al., 2016), both most widely used in the life sciences. Both of these systems allow existing software tools to be 'wrapped' into workflow components that define their required input parameters and data and output data types. The researcher can then connect the output of one tool with the input of another to construct a workflow to perform an analysis.

Fig. 1 shows a sample Taverna workflow that uses tools for searching text and finding and counting named entities to derive a set of results for disease mentions in the texts. Each of the tools in this case is either a Java beanshell executable or another composite workflow, but the researcher doesn't need to know how to execute these or pass the output of one tool to the next; this is managed by the workflow engine. The workflow in the figure is one that was shared via the myexperiment.org website which is a service for sharing workflows that can be downloaded and run on an instance of the relevant workflow engine. The main target of myexperiment.org is Taverna, although Galaxy workflows can also be shared on the site.

While Taverna provides a desktop workflow construction tool, Galaxy is a web based tool and so does not require any installation by the researcher. Galaxy also supports a more exploratory mode of work where tools can be applied one-by-one and later built up into a complete workflow. Galaxy servers are typically installed in an institution or research group and host tools relevant to the analyses carried out there. The Galaxy server also hosts published workflows and intermediate data that can be referenced in publications. Galaxy supports job execution either on a single host or via a number of job-runners on cloud-based compute clusters.

[5] https://figshare.com
[6] http://www.datadryad.org/
[7] https://zenodo.org/
[8] https://osf.io/tvyxz/wiki/home/
[9] https://guides.github.com/activities/citable-code/

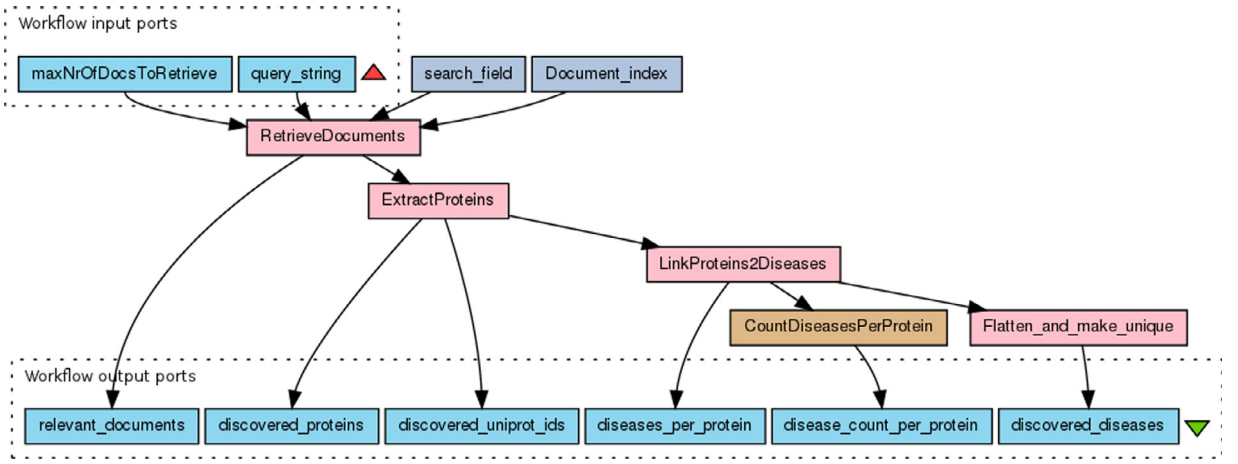

Fig. 1. A sample Taverna workflow describing a data mining task for identifying disease mentions in medical literature. Shared on myexperiment.org by Marco Roos (http://www.myexperiment.org/workflows/72.html).

Galaxy is a widely used tool in the life sciences, and particularly in bioinformatics and the analysis of DNA sequences. Workflows are cited as part of the publication process, for example Carissimo et al. (2016) cites a workflow[10] and also refers to intermediate data and analyses on Figshare.

Weblicht (Hinrichs et al., 2010) is a web based workflow tool built specifically to support language processing pipelines as part of the CLARIN-D project. As a web based workflow tool, Weblicht shares some of the features of Galaxy but in a more constrained framework. Weblicht workflows are simple chains of tools run one after the other; when composing a chain, the user is offered only those tools that are compatible with the output of the previous step. Tools are implemented as web services that are called by the tool execution engine. A number of well-known language processing tools have been included as services in Weblicht.

2.5. Summary

Following a lot of activity around the importance of reproducibility in science there are now a range of services and systems to support sharing of data and workflows. These are now starting to be used in some areas of science with data and software citation being recognised as important components of the publication of research results.

3. Language data repositories

This section provides an overview of some speech and language data repositories currently available to speech and language researchers: the CLARIN Virtual Language Observatory[11], META-SHARE[12], BAS[13], PARADISEC[14], LDC[15], ELRA[16] and CAMOMILE[17]. The main aspects we looked at were how they allow data to be shared, searched and accessed. The information presented below was collected through exploring the different web sites, trying to access the data, and through the available documentation.

[10] https://mississippi.snv.jussieu.fr/u/drosofff/p/anopheles-viruses
[11] https://vlo.clarin.eu
[12] http://www.meta-net.eu
[13] http://hdl.handle.net/11858/00-1779-0000-000C-DAAF-B
[14] http://www.paradisec.org.au
[15] https://www.ldc.upenn.edu
[16] http://www.elra.info/en
[17] https://camomile.limsi.fr

3.1. Access permissions

In this section we review the degree of openness and ease of access to data in each repository.

The CLARIN Virtual Language Observatory is open to institutions and also to individuals as guest users, with no access fee. Access to some corpora requires a fee.

For META-Net/META-Share, the preference is for institutional access from (European) partner institutions. Individuals from institutions not affiliated to META-Net can become members but this seems to create some difficulties in logging across different nodes and requesting access to datasets. Access is free of charge. Some data is cross-listed with ELRA and LDC (e.g. the 2006 CoNLL Shared Task – Arabic & Czech).

PARADISEC (the Pacific and Regional Archive for Digital Sources in Endangered Cultures) is open to institutions or individuals and provides data free of charge. In spite of its name, there are no regional restrictions on the data, but it is intended for material that is endangered in some way.

BAS (Bayerisches Archiv für Sprachsignale) was created for speech resources of contemporary spoken German and is now a licensed CLARIN centre, so can be accessed under the same conditions as CLARIN. It also provides tools for the processing of digitized speech (e.g. MAUS).

The LDC (Linguistic data Consortium) and ELRA (European Language Resources Association) both require access fees, although some corpora are available free of charge (e.g. the 2006 CoNLL Shared task data, which is cross-listed under META-SHARE). Access to both LDC and ELRA is restricted to organisations. The CAMOMILE (Collaborative Annotation of multi-MOdal, multI-Lingual and multi-mEdia documents) project gives access to some audio-visual data and tools and there is no access fee, but it is restricted to the project partner institutions and collaborators.

3.2. Metadata search

CLARIN allows faceted browsing (‘language’, ‘resource type’) of the catalogue, and simple textual queries (via Weblicht) or advanced queries of individual documents. The latter two return direct links to the matching documents.

META-SHARE allows faceted browsing of the catalogue, e.g. ‘monolingual, French, corpus, audio’ (which returns 53 results), but not of the corpora themselves.

BAS allows metadata searching through CMDI and Dublin Core via SQLite. The corpora can also be browsed at OLAC. An interesting feature is that the repository search function allows the user to define and download cross-corpora datasets.

In the LDC archive, in addition to the catalogue search from the public website with general facets such as ‘language’ or ‘corpus name’, an authenticated user can search metadata through the on-line catalogue with corpus-specific fields such as ‘genre’ (e.g. ‘humour’ or ‘reportage’) for the Brown Corpus, or ‘topic’ or ‘sex’ for the Switchboard corpus.

PARADISEC content metadata from the catalogue can be searched freely with unrestricted access through OLAC[18], ANDS[19] or the LINGUIST LIST gateway[20].

3.3. Access to individual files

META-SHARE, LDC and ELRA require users to purchase or download a complete corpus, not individual files.

BAS allows download of single files as examples for some corpora (e.g. Siemens 1000). Otherwise, the whole CD or set of CDs must be purchased, or the whole on-line corpus must be downloaded (e.g. VerbMobil). Via the BAS repository[21] academic users have access to individual files and can download single recording sessions or corpus packages.

[18] http://www.language-archives.org
[19] https://researchdata.ands.org.au/paradisec-collection
[20] http://linguistlist.org/forms/langs/find-a-language-or-family.cfm
[21] http://hdl.handle.net/11858/00-1779-0000-0006-BF00-E

PARADISEC not only allows access to single files but to items comprising several files (e.g. an audio-video recording of a meeting or dance performance).

The CLARIN VLO allows download of single files for some collections.

CAMOMILE allows access to individual files and provides a web API to support writing scripts to access data.

3.4. Sharing data

ELRA identifies the Linguistic Resources it will re-distribute. Although researchers can alert ELRA to resources they would like to see made available, ELRA does not permit researchers to deposit data.

CLARIN provides depositing services in several of the centres (Austria, Czech Republic, Denmark, Germany, Netherlands, Norway, Poland, Slovenia, USA). The resources and metadata are then integrated into the CLARIN infrastructure and provided with a persistent identifier.

META-SHARE allows upload of resources by members. The resource must be described with appropriate metadata and users can choose a license type (Creative Commons or META-SHARE No Redistribution).

BAS is one of the CLARIN centres which allows researchers to ingest their own data; it offers validation of the speech resource being deposited.

PARADISEC not only allows researchers to deposit their data with PARADISEC, the interface to the PARADISEC Catalog is designed for researchers to add to their own collections while still collecting the data. A minimum set of metadata is required and depositors can specify conditions of use for each collection and even each item.

LDC allows LDC members to contribute to the LDC catalogue.

CAMOMILE allows its project partners to add data to the resources to be annotated.

3.5. Summary

In summary, four of these repositories provide both free and easy access to existing data and allow researchers to share their data. Five repositories provide metadata search. Four allow access to data within a collection one file at a time. Except for ELRA, the repositories allow at least partner users to upload their own data for sharing, and most provide some level of validation for the new collections to be added. CLARIN, META-SHARE, BAS and LDC offer software tools, while CAMOMILE is a platform for annotation, which can be considered a form of tool.

All of the repositories reviewed provide a long-term storage facility for language data with facilities to search for data based on various metadata properties. In all cases, the repositories are designed to accept submissions of complete collections with the exception of CAMOMILE which offers a platform to help in the annotation of language data via an API. In most cases, downloads from the repository are of the entire dataset, although some repositories allow individual files to be requested. All of these repositories support reproducibility by offering a secure home for data collections that authors have decided to make widely available.

4. The Alveo Virtual Laboratory

4.1. Motivation

Alveo[22] is a *Virtual Laboratory* to support research in language and speech. It combines a data repository with a web API and interfaces to a workflow platform and seeks to address some of the issues raised in this paper so far.

The design of Alveo derives from the earlier DADA system (Cassidy, 2010) which had a focus on the management of annotations over linguistic data; however, the goals for Alveo are broader with the overarching goal of providing an accessible workbench that will support sharing of data and simplify the application of complex tools as part of a research workflow.

One motivating insight is that many of the tools developed in some disciplines are relevant to others where researchers may not have the technical knowledge to acquire and apply them in their work. As an example, consider the *speaker diarization* technology developed by speech technology researchers that can identify different talkers in

[22] https://alveo.edu.au/

a recording. In the context of social history research, this might be useful as an aid to the transcription of audio interview recordings; however, few historians would know about this technology let alone be able to apply it to their data. Our goal with Alveo is to provide a way for technology researchers to make their tools available in such a way as to be discoverable by those in other disciplines.

In the same way, it is likely that data collected for research in one discipline will be of use in other contexts. An example would be the use of oral history data in studies of dialogue structure or as the input data for training speech technology tools. Bringing a wide range of data into the same technical framework provides the opportunity for serendipitous discovery of new ways of looking at a wide range of research questions.

Alveo is a data repository for language and speech data, but unlike other similar systems it makes data available down to the level of individual documents and audio files. While data is available for download in bulk, Alveo also allows the researcher to identify the subset of the data that is of interest for a study and work only with that data. Importantly, these data subsets must be reproducible and citable.

4.2. Data structures

The data model that we have developed for the storage of language resources is built around the concept of an *item* which corresponds (loosely) to a record of a single communication event. Arbitrary metadata can be associated with an item, for example describing the title and creation date of a text or the speaker and language of an audio recording. An item is often associated with a single text, audio or video resource but could include a number of resources, for example the different channels of audio recording or an audio recording and associated textual transcript. In the system these are denoted as *documents* and can have associated metadata of their own (e.g. describing the audio channel or provenance of the transcription). One of these documents is denoted as *indexable*, meaning that it will be indexed for full-text search (and so it would normally be a plain text version of the item content).

A group of items is known as a *collection*, avoiding the somewhat loaded term 'corpus' which is often reserved for carefully curated collections of language. A *collection* is a set of items and has associated metadata describing the collection as a whole. Collections may correspond to formally curated corpora such as the Australian Corpus of English (ACE[23]) or to more informal collections like a sample of audio excerpts from Disney Pixar movies expressing different emotions (pixar[24]).

Each collection is described by metadata using a standard Dublin Core description, however the main metadata stored is at the level of *items* within a collection. These are described using a mixed metadata vocabulary based on the descriptions that are provided to us by the original data owners. While the range of metadata fields available varies significantly across the collections, we have mapped the fields to common names where possible using a mixture of Dublin Core, OLAC and custom namespaces. The result is that there are only a small number of fields that will be available for all items in the system but that common names are used where possible to facilitate searching across collections. Fig. 2 shows an extract from a metadata record and illustrates the combination of well-known metadata properties and those specific to a single collection.

An important part of many studies is the selection of a subset of data to focus on for a particular question. This may be a partition of the data into training and test sets or the selection of a group of speakers or words relevant to a question. Alveo provides a data structure called an *item list* to reflect this practice. An item list is a simple list of items from the Alveo repository. It can contain items from one or more collections and can be built up through one or more queries over the item metadata. Importantly, item lists can be made public and shared via a URL so that others can access the same collection of data.

While audio, video and text resources form the primary data types for communication research, annotations are an important value-add that turn a simple collection of data into a useable resource. Annotations can be generated manually or through some automated processing pipeline. They can be stored in Alveo as documents within an item alongside the primary data; for example, a TextGrid file generated by forced alignment of a transcript with an audio recording. In addition to this, Alveo supports an *annotation store* that is able to store annotations in a graph based format and associate them with items in the repository. Annotations stored in this way can be queried either at the level of the item or at the level of the collection. Behind the scenes, annotations are stored in an RDF Triple Store

[23] https://app.alveo.edu.au/catalog/ace
[24] https://app.alveo.edu.au/catalog/pixar

```
"dc:created": "1851",
"dc:identifier": "3-032",
"dc:isPartOf": "cooee",
"ausnc:discourse_type": "narrative",
"olac:language": "eng",
"cooee:register": "Public Written",
"cooee:texttype": "Memoirs",
```


Fig. 2. Sample metadata record for an item from the COOEE collection showing some common properties (`dc:created`) as well as collection specific properties (`cooee:register`).

using an RDF format based on the ISO-LAF data model (Cassidy, 2010). Annotations and metadata can be queried either via the API, when a JSON-LD representation is made available, or via a SPARQL query interface.

4.3. Authentication and authorisation

By nature, many of the collections that we are dealing with are not able to be made open or released under a liberal licence. In some cases this is due to the original terms under which the data were collected, in others because of cultural or privacy issues relating to the people who were recorded. To deal with this, all access to data on the platform is mediated by an authorisation layer which checks which licence agreements apply to a collection and whether a user has agreed to those terms or been granted access to that collection by the data owner. On registering with Alveo, a user is able to review the collections that are held and the licence terms under which they are available; if the user agrees to the licence terms they can be granted access to that collection. In some cases, access is mediated through the data owner or a delegate; this allows us to handle more complex licence arrangements such as when a fee is paid to a third party for access or where membership of a particular institution is required.

The licences used with each collection are those provided by the collection owner. Many of the older collections we hold have existing licences. Any new collections are encouraged to adopt an open access licence such as one of the Creative Commons suite. Alveo is agnostic as to the licence used for collections that are deposited as long as they are compatible with the basic model where users agree to licence terms via an online link. If data owners wish to enforce stricter conditions they would have to mediate permissions on Alveo and manage permissions themselves.

All access to data is mediated via this authorisation layer including search via the website and all access via the web API. To use the API, a user downloads an authentication token which allows a script to act as a proxy for the user, gaining access to data that they are permitted to access.

4.4. Web API

All data and services of Alveo are made available via a RESTful web API (Cassidy et al., 2014). All entities in the system (collections, items, documents, annotations, etc.) are identified via a URI and, following the principles of Linked Data, that URI resolves to a representation of that entity. HTTP content negotiation is used to determine whether to return an HTML or JSON representation for any URL.

Use of the API requires authentication. This is achieved by a token that is unique to each user and can be downloaded from the website. The token is sent along with every HTTP request using the X-API-KEY header. A user can invalidate their API token via the website if it is accidentally published and a new key can be generated. This mechanism allows a client application to operate on behalf of a user via a relatively simple and reasonably secure mechanism. We have recently extended the authentication system to support OAuth2 which is particularly useful if the client application is another website.

The API supports both reading and writing data to the Alveo repository. New collections can be created and within them, new items added with metadata and documents. Once items are in place, new annotations for the documents in an item can be added using the same JSON representation that is used for download of annotations. The API can also be used to create new item lists and add items to existing lists. Using these facilities, we are able to write applications to allow users to easily add their data collections to the repository.

```
# create an Alveo client interface
client = pyalveo.Client(use_cache=False)

# get a list of item URLs via a metadata query
query = "collection_name:austalk␣AND␣speaker:1_1308"
items = client.search_metadata(query)

# for each item url, get item metadata
for itemurl in items:
    item = client.get_item(itemurl)
    meta = item.metadata()
    speakerid = meta['alveo:metadata']['olac:speaker']

    # create a subdirectory based on speaker id
    subdir = os.path.join(outputdir, speakerid)
    if not os.path.exists(subdir):
        os.makedirs(subdir)

    # download the documents attached to this item
    for doc in item.get_documents():
        doc.download_content(dir_path=subdir)
```

Fig. 3. A sample script fragment that uses the `pyalveo` library to access data on Alveo. The `client` object establishes a connection to the Alveo server and various methods are used to get item lists, items and documents.

To make using the API easier, we have implemented libraries in a number of programming languages that provide an interface to reading and writing data on Alveo. The most fully developed of these is the `pyalveo` module for the Python language[25], but modules are also available for R, Java and Go[26]. These modules allow scripts to be written to run queries, download data and annotations and also to add new data to the Alveo repository. Fig. 3 shows a small extract from a script that uses the `pyalveo` library to query and download data from Alveo.

4.5. Galaxy workflows for language

While the core of the Alveo platform is the data repository and API, we are also developing tools to support workflows for language and speech data on the Galaxy workflow engine. Galaxy was originally developed as a platform for bioinformatics research and the majority of tools available are related to this domain. However, the Galaxy platform is agnostic to the kind of data that is being analysed and with the support of the core Galaxy developers there are now a number of groups around the world developing tools for language processing.

A Galaxy tool is simply an executable script or application that reads data from one or more files and writes results to one or more files. Galaxy recognises a large number of file formats and new formats can be added; tools can be configured to accept a particular kind of file as input with a goal that only valid workflows can be generated. To convert an existing executable into a Galaxy tool, a small XML configuration file is written that sets out the input and output parameters. This file is used by the system to generate a user interface for invocation of the tool. As an example, Fig. 4 shows the interface for a tool based on the MAUS forced alignment system.

A number of Galaxy tools for language are in development with some deployed on our server at http://galaxy.alveo.edu.au. These include wrappers for speech processing tools (wrassp[27]), plotting libraries and the MAUS forced alignment tool Strunk et al. (2014). Other tools are being developed for text processing including those based on the NLTK Python library (Bird et al., 2009) and the Stanford NLP toolkit (Manning et al., 2014).

In addition to our own work with Galaxy we are aware of the work of the Language Application Grid (LAPPS) project (Ide et al., 2014) who have converted a number of web-services based NLP tools for use in Galaxy, and the work of Lapponi et al. (2014) building tools that run at scale on a high performance computing cluster. In both cases,

[25] https://github.com/Alveo/pyalveo
[26] https://github.com/Alveo
[27] https://github.com/IPS-LMU/wrassp

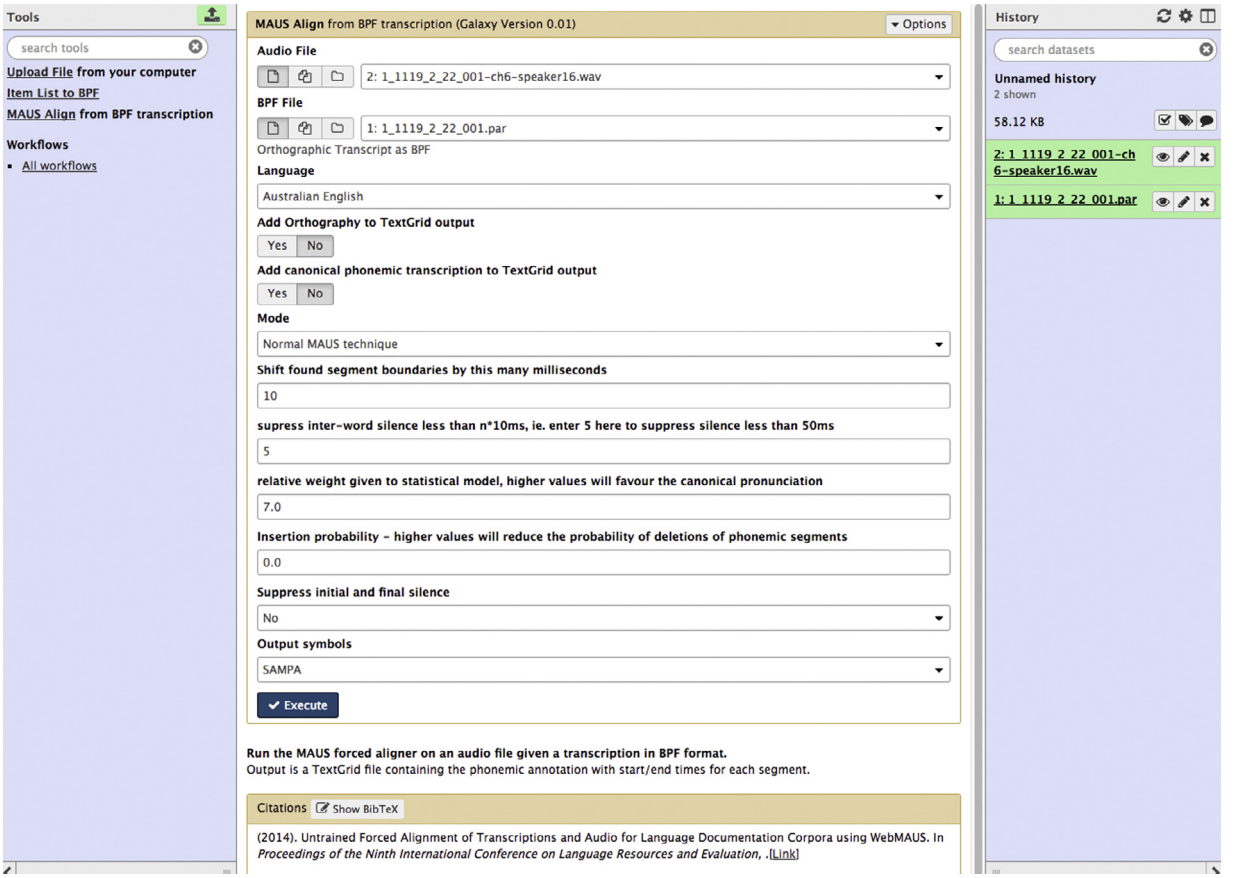

Fig. 4. Screenshot of the Galaxy tool interface for a tool based on the MAUS forced alignment tool.

these projects are focussed on text processing pipelines, making use of existing software such as the Stanford NLP tools and MaltParser.

A user can run a Galaxy tool manually by selecting the tool and then choosing values for the various input parameters. Running the tool schedules a job on the backend server, which can be a simple script, an interface to a compute cluster or a call to a remote web service. Once the job has finished, the output is available as a named item in the user's history and can be viewed or used as the input to a subsequent tool. Once the sequence of operations is understood, a workflow can be constructed either from the history of tool executions or via the workflow editor. Fig. 5 shows a simple workflow that applies a formant tracker to audio data to extract the formant values at the mid-point of hand labelled vowels. This workflow makes use of the `tgt` Python library (Hendrik Buschmeier, 2013) to query Textgrid files and the `wrassp` R package to generate the formant tracks. This illustrates the ease with which different programming languages can be mixed in a single workflow with the user not having to be concerned about the runtime requirements of the tools, only the job that they do.

4.6. Relation to other repositories

In Section 3 we reviewed a number of data repositories that hold language data and might be seen as contemporaries of the Alveo system. This section compares Alveo under the same headings.

Alveo requires a user to be registered to access any data or search metadata of collections and data. Registration is available to anyone and the system supports federated login for Australian researchers via the Australian Access Federation. Before a user is able to access any data they must review and agree to licence terms for each collection

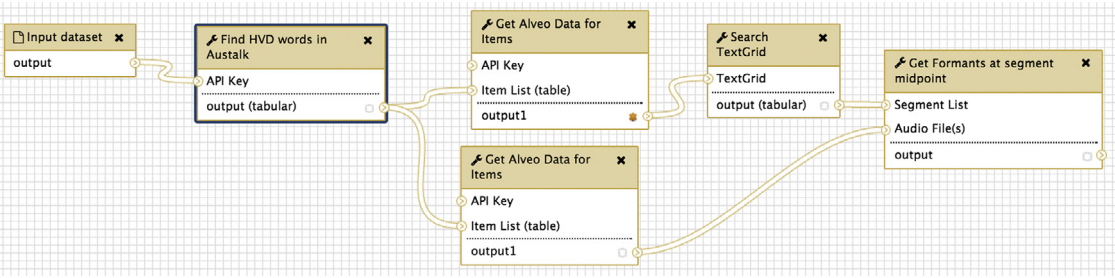

Fig. 5. A Galaxy workflow to generate a table of formant values for hand-labelled vowel data from the Austalk collection.

and in some cases be granted permission by the data owner. Alveo has no provision for charging fees for access to data.

All of the archives reviewed provide a metadata search facility at the level of collections and some support searching the metadata attached to individual files and documents inside the collections. Alveo supports both of these and emphasises the metadata on items more than that on collections; the intention is to encourage users to search across collections using relevant metadata fields if they are present. To further support this mode of work, Alveo provides the *item list* mechanism that allows researchers to save and share their own sub-collections that might contain items from many parent collections.

A significant feature of the Alveo system is the API access to individual items and documents that allow tools to process data directly from the archive. No other archive reviewed provides this level of access, although some do allow individual files to be downloaded via the web.

Alveo has been designed to accommodate new contributions from researchers either as new collections or new annotations on existing collections. The API supports adding new items to collections, adding new documents to existing items and adding annotations. We are working on user-level tools to make these contributions easier for researchers. All of the repositories reviewed, except the PARADISEC archive, support a model of contribution of completed collections only.

5. Discussion

This paper has presented a summary of the current state of practice in speech and language research relevant to the reproducibility of research findings. The overall picture in the field is positive in that most of the papers in the small sample reviewed paid attention to the need to properly describe the data and methods used in the study and at least attempted to make new datasets available in many cases. There is an understanding that being able to reproduce the work described in a paper is an important part of making it a useful publication. However, we found that the way that authors referred to data and the way that they made new data available showed room for improvement.

The Alveo Virtual Laboratory is a web based repository for language data that is designed to support the research workflow in a way that enhances the reproducibility of results.

As a repository for language data, Alveo acts as a source of well-known datasets that can be used in research studies but also provides a home for new datasets developed as part of the study. Authentication and access control measures ensure that data in the repository is only shared as widely as desired by the author. Datasets can be referenced by URL giving a reader direct access to the source data for a study; in the case that the data is protected by access control policies, the reader will be directed to the author for permission to use the data. In this way, Alveo provides a way for researchers to store data in a robust repository while they generate their results and transition to a published dataset with little effort when appropriate.

If Alveo were just a data repository it would not be a significant innovation. We reviewed a number of existing repositories for language data and all provide stable stores of data that authors wish to share with others. All of these repositories assume that the data owner will contribute a dataset at some point after the work is complete in order to make it available to others. In contrast, the design of Alveo is intended to encourage researchers to add their data early in the research cycle and use the facilities of the platform as part of their research workflow. To make this an

attractive proposition, Alveo provides an API that supports access to individual data files within a collection and a set of tools that automate parts of the research workflow.

Software tools are a critical part of any work with digital language resources and papers often describe new methods embodied in new software as part of the contribution of a study. In common with many other computer science disciplines, the problem of reproducibility comes down to making software code available and taking steps to ensure that colleagues can build and execute it. Other disciplines in language and speech are similarly reliant on software tools but do not have the expertise to write them; they rely on a range of tools to compute features, train models and generate analytics often driven by scripting languages such as R. One goal of the Alveo platform is to integrate these tools with the API so that they can operate directly on data from Alveo. Further to this, we explore the Galaxy workflow engine to provide a web based interface for running complex tools and constructing automated workflows. The goal of all of this work is to make it easier for researchers to cite a version of their workflow that is detailed enough to be fully reproducible. In this we seek to learn from the work that has been done in other disciplines and bring their best practices to the speech and language domain.

Alveo is by no means the final answer to reproducibility in speech and language research. As with most systems of this type, it will be successful in some areas and need adaptation in others. The success of the system will depend on the degree to which it can help researchers generate their research results. If it makes it harder to make progress, it will not be used. The goal is to make it easier to build research workflows that are more reproducible than before.

6. Future work

Alveo is an ongoing project and development continues on specific parts of the platform. In writing this review, a number of next steps are apparent that might improve the repeatability of results generated with the Alveo platform.

Currently Alveo resources (collections, items, item lists) are referenced via a URL that resolves directly to the resource on the Alveo server. If the user does not have permission to access the resource then they are directed to the relevant licence agreement or to contact the data owner. While URLs in Alveo should be more robust than those in a research group website, they are still subject to possible change in time as funding sources dry up or change. A more robust method of referring to digital resources would be the Handle System, perhaps via a Digital Object Identifier (DOI). The Handle system provides a resolver service for digital identifiers which can be updated if the back-end storage address changes in future. A DOI is also a more acceptable method of citing a digital resource in publications and may allow authors to be given citation credit for their resources separately from their papers. The Alveo project should consider allowing data owners and researchers to easily generate a Handle or DOI reference for a collection or item list.

Following the example of some of the general purpose research workbenches mentioned above (Figshare, Dryad, Zenodo), Alveo could provide a richer workspace to share more of the intermediate results in a study as well as the item list and workflows that it currently supports. To achieve this, it may be possible to integrate Alveo resources with an existing research workbench. This may be preferable to re-implementing this kind of functionality in a discipline specific workspace linked only to the Alveo system.

Alveo does not yet support any kind of version management for collections. A new version of a collection in Alveo would have to be stored as a separate new collection in the system. Clearly version management is an important feature for a modern repository and we have some ideas of how to address this. Version management is important for the original research team but also for researchers who seek to add to or improve a resource that has been in use for some time. One model to look to is the *forking* model used in software repositories like Github and Bitbucket. A researcher could create a copy or fork of a collection to add new material or create changes. Behind the scenes the system could keep track of what is new or different to avoid keeping a complete new copy of the data. This would be important for understanding the relation between the new and old collections as well as for storage efficiency. Another idea would be to follow an *additive* model for new contributions or changes; a researcher could submit a collection of files which when merged with the original created the updated collection. These contributions could be given a similar status to collections with citable DOIs to allow attribution to the authors of the changes and the original work.

An important consideration for a platform like Alveo is the longevity and sustainability of the repository for research data. The development of Alveo is supported by thirteen Universities around Australia and they remain interested in the long-term maintenance of the service it provides; in particular the home institutions of the authors

are supportive and have to some extent integrated hosting of Alveo into their broader e-research support. We are working with the research community in Australia and internationally to try to ensure that the technology behind Alveo can be maintained and become an active part of the research workflow. Based on past experience, Alveo will be a sustainable platform if there are many researchers interested in seeing it work and a core of active developers able to maintain the system. Our goal is to grow such a community around Alveo as a research platform.

Acknowledgments

The Alveo project was funded by a grant from the The National eResearch Collaboration Tools and Resources project (NeCTAR); NeCTAR is an Australian Government project conducted as part of the Super Science initiative and financed by the Education Investment Fund. The project also acknowledges funding from 13 Australian Universities for the ongoing work on the Alveo platform.

References

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., Goecks, J., 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. gkw343. doi: 10.1093/nar/gkw343.

Bayer, A.O., Riccardi, G., 2016. Semantic language models with deep neural networks. Comput. Speech Lang. 40, 1–22. doi: 10.1016/j.csl.2016.04.001.

Berners-Lee, T., 1998. Hypertext Style: Cool URIs Don't Change. Retrieved November URL http://www.w3.org/Provider/Style/URI.html.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

Calvo, H., Méndez, O., Moreno-Armendáriz, M.A., 2016. Integrated concept blending with vector space models. Comput. Speech Lang. 40, 79–96. doi: 10.1016/j.csl.2016.01.004.

Carissimo, G., Eiglmeier, K., Reveillaud, J., Holm, I., Diallo, M., Diallo, D., Vantaux, A., Kim, S., Ménard, D., Siv, S., Belda, E., Bischoff, E., Antoniewski, C., Vernick, K.D., 2016. Identification and characterization of two novel RNA viruses from anopheles Gambiae species complex mosquitoes. PloS One 11 (5), e0153881. doi: 10.1371/journal.pone.0153881.

Cassidy, S., 2010. An RDF Realisation of LAF in the DADA Annotation Server. In: Proceedings of ISA-5. Hong Kong.

Cassidy, S., Estival, D., Jones, T., Burnham, D., Burghold, J., 2014. The Alveo virtual laboratory: a web based repository API. In: Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland.

Cohen, K., Xia, J., Roeder, C., Hunter, L., Graliński, F., Jaworski, R., Borchmann, Ł., Wierzchoń, P., Francopoulo, G., Mariani, J., Paroubek, P., 2016. Reproducibility in natural language processing: A case study of two r libraries for mining PubMed/MEDLINE. In: Proceedings of the 4REAL Workshop. URL http://4real.di.fc.ul.pt/.

Collberg, B.Y.C., Proebsting, T.A., En, W.H., Collberg, C., Proebsting, T.A., 2016. Repeatability in Computer Systems. Commun. ACM 59 (3), 62–69. doi: 10.1145/2812803.

Ghaemmaghami, H., Dean, D., Sridharan, S., van Leeuwen, D.A., 2016. A study of speaker clustering for speaker attribution in large telephone conversation datasets. Comput. Speech Lang. 40, 23–45. doi: 10.1016/j.csl.2016.03.005.

Gonzalez, J.A., Cheah, L.A., Gilbert, J.M., Bai, J., Ell, S.R., Green, P.D., Moore, R.K., 2016. A silent speech system based on permanent magnet articulography and direct synthesis. Comput. Speech Lang. 39, 67–87. doi: 10.1016/j.csl.2016.02.002.

Hendrik Buschmeier, M.W., 2013. TextGridTools: A TextGrid Processing and Analysis Toolkit for Python. In: Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013), pp. 152–157.

Hinrichs, E.W., Hinrichs, M., Zastrow, T., 2010. Weblicht: Web-based lrt services for german. In: Proceedings of the ACL 2010 System Demonstrations, pp. 25–29.

Ide, N., Nyberg, E., Suderman, K., Wang, D., 2014. The Language Application Grid. LREC 22–30. doi: 10.1007/978-3-319-31468-6_4.

Kelly, C.W., Chase, L.J., Tucker, R.K., 1979. Replication in Experimental Communication Research: an Analysis. Human Commun. Res. 5 (4), 338–342. doi: 10.1111/j.1468−2958.1979.tb00646.x.

Kidwell, M.C., Lazarević, L.B., Baranski, E., Hardwicke, T.E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T.M., Fiedler, S., Nosek, B.A., 2016. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. PLOS Biol. 14 (5), e1002456. doi: 10.1371/journal.pbio.1002456.

Lapponi, E., Velldal, E., Oepen, S., Lain Knudsen, R., 2014. Off-Road LAF: encoding and processing annotations in NLP workflows. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Piperidis, J.O.S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 4578–4584. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/978_Paper.pdf.

Lopez-Moreno, I., Gonzalez-Dominguez, J., Martinez, D., Plchot, O., Gonzalez-Rodriguez, J., Moreno, P.J., 2016. On the use of deep feedforward neural networks for automatic language identification. Comput. Speech Lang. 40, 46–59. doi: 10.1016/j.csl.2016.03.001.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D., 2014. The Stanford CoreNLP natural Language Processing toolkit. Association for Computational Linguistics (ACL) System Demonstrations, pp. 55–60.

Muma, J.R., 1993. The need for replication. J. Speech Hearing Res. 36 (5), 927–930. doi: 10.1044/jshr.3605.927.

Open Science Collaboration, 2015. Estimating The Reproducibility of Psychological Science. Science 349 (6251). doi: 10.1126/science.aac4716 arXiv:1011.1669v3. URL http://www.ncbi.nlm.nih.gov/pubmed/26315443

Prasad, A., Ghosh, P.K., 2016. Information theoretic optimal vocal tract region selection from real time magnetic resonance images for broad phonetic class recognition. Comput. Speech Lang. 39, 108–128. doi: 10.1016/j.csl.2016.03.003.

Qin, Z., Cong, Y., Wan, T., 2016. Topic modeling of Chinese language beyond a bag-of-words. Comput. Speech Lang. 40, 60–78. doi: 10.1016/j.csl.2016.03.004.

Reese, S., Boleda, G., Cuadros, M., Padró, L., Rigau, G., 2010. Wikicorpus: a word-sense disambiguated multilingual wikipedia corpus. In: Proceedings of 7th Language Resources and Evaluation Conference (LREC'10), pp. 1418–1421.

Spinu, L., Lilley, J., 2010. Classification of gender based on cepstral coefficients and spectral moments. J. Acoust. Soc. Am. 127 (3), 1855. doi: 10.1016/j.wocn.2016.05.002.

Strunk, J., Schiel, F., Seifart, F., 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using webmaus. In: Calzolari, N., Choukri, K., Declerck, T., Doan, M.U., Maegaard, B., J, M., Odijk, J., St, P. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation. European Language Resources Association.

Wilson, G., 2006. Software carpentry: Getting scientists to write better code by making them more productive. 8 (6), 66–69. doi: 10.1109/MCSE.2006.122.

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., Nieva de la Hidalga, A., Balcazar Vargas, M.P., Sufi, S., Goble, C., 2013. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Res. 41 (Web Server issue). doi: 10.1093/nar/gkt328.

Zamani, H., Faili, H., Shakery, A., 2016. Sentence alignment using local and global information. Comput. Speech Lang. 39, 88–107. doi: 10.1016/j.csl.2016.03.002.

Zhang, W., Clark, R.A., Wang, Y., Li, W., 2016. Unsupervised language identification based on Latent Dirichlet Allocation. Comput. Speech Lang. 39, 47–66. doi: 10.1016/j.csl.2016.02.001.