



LAT NEWS

November 2011

Table of Contents

November 2011	1
<u>Statistical Language Models for Alternative Sequence Selection</u>	2
<u>The Language Archive officially launched</u>	2
<u>The Endangered Language Catalog (ELCat)</u>	4
<u>West Ambrym in the Humboldt-Box</u>	4
<u>Summary of the 2011 CLARA Summer School</u>	7
<u>Upcoming Projects with TLA participation</u>	10
<u>Introduction of a new transcription mode in ELAN 4.1.0</u>	12
<u>Semantic interoperability of linguistic resources now and in the future</u>	14

LAT News is made by the Language Archive of the Max Planck Institute for Psycholinguistics.

This is a PDF version of content originally published on the web site <http://www.lat-mpi.eu/latnews/> between April and November 2011.

Editor of this issue: Alexander König <Alexander.Koenig@mpi.nl>

Contact: latnews@mpi.nl



Statistical Language Models for Alternative Sequence Selection

by Herman Stehouwer

Is there a need to limit certain aspects of statistical language models?

Is it necessary to pre-limit the size of the n-gram?

Is it useful to use linguistic annotation, within alternative sequence selection tasks?

According to a new study by Herman Stehouwer, the size of the n-gram can be completely flexible depending on the situation. The study also finds that the addition of certain linguistic annotations, specifically part-of-speech annotations and dependency-parses, did not aid the model in making decisions.

The study compares the ability of a language model to select the correct alternative from sets of alternatives in hundreds of experiments. These experiments were performed for three different alternative sequence selection tasks, for four different annotations (and also for no annotation), and for four different ways to combine the annotation with the text. The results of the study have been used to write the thesis "Statistical Language Models for Alternative Sequence Selection". This thesis will be defended on the 7th of December at 18:00 in the Aula of Tilburg University.

Coinciding with the defense a colloquium on language modeling is organized with invited talks by Colin de la Higuera, Louis ten Bosch, and Antal van den Bosch. For more information on the colloquium you can send an e-mail to herman.stehouwer [at] mpi.nl or look at its [website](#).

The Language Archive officially launched

by Sebastian Drude

Tuesday, the 11th of October 2011, the new unit of the Max-Planck-Institute for Psycholinguistics "The Language Archive" (TLA) has been officially launched in a public event with more than 150 guests and speeches from eminent representatives from Germany and the Netherlands.

Many more showed up than expected: there were even not enough seats for all



guests at the launching of TLA in the Headquarters of the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) at the Gendarmenmarkt in the center of Berlin. The BBAW is one of the three supporting institutions of TLA, together with the Dutch Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) and the German Max-Planck-Gesellschaft (MPG).

The guests were presented with coffee and snacks, but before and above all with much content: five eminent representatives of the major stakeholders of the new unit gave fascinating talks discussing different topics, all related to the ongoing and future activities of TLA. These were on the one hand the respective representatives of the three supporting institutions: Wolfgang Klein for the MPG, Angelika Storrer for the BBAW, and Theo Mulder for the KNAW. On the other hand, Wilhelm Krull represented the Volkswagenstiftung, the funding agency that supports the programme "Documentation of Endangered Languages" (DOBES) since 2000, which in turn was represented by Nikolaus P. Himmelmann. The DOBES archive is in many respects the core of the archive hosted by TLA. After the talks, Paul Trilsbeek provided a look into the archive itself.

The full program and topics of the speeches

Begrüßung und Zielstellung für das Spracharchiv Prof. Dr. Wolfgang Klein *Direktor am Max Planck Institut für Psycholinguistik*

Sprachforschung und Sprachdokumentation im digitalen Zeitalter Prof. Dr. Angelika Storrer *Zentrum Sprache der BBAW*

E-science: a major challenge for the humanities Prof. Dr. Theo Mulder *Forschungsdirektor der KNAW*

Dokumentation bedrohter Sprachen – eine Aufgabe für Wissenschaft und Gesellschaft Dr. Wilhelm Krull *Generalsekretär der VolkswagenStiftung*

Wie die Sprachwissenschaft zur Empirie fand (und findet) Prof. Dr. Nikolaus P. Himmelmann *Universität Köln*

Blick ins Archiv (interactive presentation)

The TLA Opening in the media:

[Official press release of the MPI-PL](#)

[Online reports on the TLA opening](#)

[Süddeutsche Zeitung](#)

[Der Tagesspiegel](#)

[Märkische Allgemeine](#)

[Frankfurter Rundschau \(article\)](#) (the same article at [Berliner Zeitung](#))



[Frankfurter Rundschau \(interview with Wolfgang Klein\)](#) (the same at [Berliner Zeitung](#))

ix

The TLA opening in the paper press

[Süddeutsche Zeitung \(PDF\)](#)

[Berliner Zeitung \(PDF\)](#)

[Frankfurter Rundschau \(PDF\)](#)

[The Endangered Language Catalog \(ELCat\)](#)

by Maddalena Tacchetti

The Endangered Languages Catalog project (ELCat) has been recently approved and made feasible by a donor's participation that will eliminate the burden of the cost of most of the technology. The ELCat is a collaborative project with the University of Hawaii and the Eastern Michigan University that aims to provide accurate, up-to-date information on the endangered languages of the world as well as raising public awareness, promoting increased research on endangered languages, but also providing the communities whose languages are at risk with materials to support language preservation and revitalization activities. The project, with [Lyle Campbell](#), Anthony Aristar, and Helen Aristar-Dry at the head of an international team of Regional Directors and graduate student researchers, will be carried on for about three years (from August 2011 until January 2015).

Already from December on, a preliminary ELCat website will be launched with mechanisms for feedback, commentary, suggestion of corrections by users and a procedure for long-term maintenance and updating of the information and the web presence.

You can read more details on the web site of the [National Science Foundation](#).

[West Ambrym in the Humboldt-Box](#)

by Lena Karvovskaya and Soraya Hosni

The DoBeS Project "Languages of Southwest Ambrym" is happy to invite you to an exhibit in the newly opened [exhibition-center Humboldt-Box](#) in the heart of Berlin. The exhibit "Sprachdokumentation auf Südwest-Ambrym" ([Flyer with more information](#)) will be open to the public from 1st of July till 31st of December 2011.

The project team members wanted the installation to present the different ways

in which culture, language and knowledge are transmitted within written (books and recordings) and oral societies (sand drawing and story telling). The highlights of the installation are **sandroings**: a unique form of art practiced in Vanuatu. An example of such a performance is shown in a short film "The Liliwi masks story" projected on the ground. The film shows an elder man drawing complex geometric figures onto the sand with a continuous one finger movement so that it will end up forming a specific picture. The drawing is followed by a story or a description. This is a sandroing performance. The Liliwi masks story has a sand drawing to illustrate the narrative.



A typical sandroing

The exhibit shows an original Sandroing left by Abel Taho as he was our guest in Berlin from Ambrym. Visitors can also try themselves to make the performance, all you need to do is to follow the instructions which a young girl on the video is giving you: Joelyne teaches German children how to draw a breadfruit. Additionally you can watch a film on the process of linguistic fieldwork at the installation. One can see how the recordings are being transcribed and translated and how a dictionary is being composed. There is also a beautiful illustration for the dictionary done by local artist Joebang Maaseng.

For those who want to see and hear more about the "Languages of Southwest

Ambrym", there is a [video channel on Youtube](#), where Soraya Hosni shares her works. At the moment it contains the film about language documentation, the video of the Liliwi sandroing performance and two films which give you instructions on how to make a sandroing yourself. The channel will be regularly updated with new films.



Visitors at the Ambrym exhibition

The project "Languages of Southwest Ambrym" is also presented to the broader public through "Science movies", the videoblog of the Volkswagen foundation. "[Wer spricht noch Daakaka?](#)" is a series of 10 shorts, filmed by Susanne Fuchs and Soraya Hosni, in which we follow them on their journey from Berlin to Ambrym. We learn about daily life in the island, from preparing meals and basic hygiene to how houses are built or marriages are celebrated. We can admire the unique volcanic landscape and tropical vegetation but we can also learn about how the "Languages of Southwest Ambrym" team conduct linguistic and ethnographic fieldwork and collaborate with local leaders, schools and children to make the best out of the research and contribute to the survival of the Ambrym language and culture for future generations.

The Project "Languages of Southwest Ambrym" has started in August 2009. It investigates three language varieties spoken on Ambrym, a volcanic island in



the northern part of Vanuatu: Daakaka, Daakiye and Dal kalaen. The goal of the project is documentation of linguistic and cultural heritage of the people of Ambrym. During extensive fieldwork sessions the team members make recordings of custom stories and cultural practices. Among others the project has created a collection of sandroings. Each drawing has been documented together with the language performance.

The team members are: Prof. Dr. Manfred Krifka, Soraya Hosni, Kilu von Prince, Dr. Susanne Fuchs and Lena Karvovskaya (student assistant). To learn more about the Project "Languages of Southwest Ambrym" visit the official websites at the MPI or at the ZAS.

Summary of the 2011 CLARA Summer School

by Przemek Lenkiewicz

The CLARA Summer School on Infrastructure Tool Development has taken place at Max Planck Institute for Psycholinguistics on 5th – 12th July.

Participants came from several institutions, including the University of Bielefeld, the Technical University of Aachen, Gießen University or Technical School of Mittelhessen. Some representatives of Max Planck staff also participated in parts of the summer school, especially those requiring less technical expertise. Altogether they have created a very inspiring and productive group that managed to carry out the tasks planned for the event and also came up with some new ideas for developing useful things, which also have been done during the summer school.

On the first day Przemek Lenkiewicz opened the summer school and introduced participants to the agenda and all extra activities. Participants were also encouraged to present themselves and their work, giving an idea about how they use ELAN and what are they hoping to learn at this event.

Later Han Sloetjes, the main developer of ELAN, has presented the annotation tool and introduced its mechanisms for creating and integrating extensions (recognizers). Some users said that although they have used ELAN for quite a long time, they were not even aware that it is possible to extend its functionality and that it is so simple. Han has spent the whole day with participants to clear out any doubts they might have. He also showed up on following days and participated in the development sessions.



Stefano Masneri with participants

Days 2-4 of the event were about signal processing techniques. Stefano Masneri of Fraunhofer HHI Berlin and Dr. Rolf Bardeli of Fraunhofer IAIS Sankt Augustin have introduced the participants to video and audio processing basics. In the afternoon hands-on sessions participants have developed some simple video/audio processing algorithms, like histogram calculations for both audio and video, color-to-greyscale conversion, image flipping, etc. But also more advanced functionality was developed, like detecting a person's hand in a video using edge detector as the base or detecting fricatives in a speech recording using thresholding.

The last two days of the summer school were led by Przemek Lenkiewicz and Eric Auer. In a brainstorming session with the participants we defined two recognizers, which are interesting for them to develop. Those included automated importing of eye-tracking data into ELAN and representing it as annotations and curves, and also a recognizer to compare two tiers based on the similarity of the annotations. Both recognizers have been successfully developed until the end of the summer school.



Eric Auer and Przemek Lenkiewicz

Since the summer school included the weekend, the group met and explored Nijmegen for a while. On Monday July 11th we also had dinner together in a nice Dutch restaurant.

Additional pictures from the event can be found on [this web page](#).

After the event participants have filled a survey and rated the summer school very well for a good content, good way to deliver it and for overall organization. Considering the good feedback, another Summer School on Infrastructure Tool Development might take place at Max Planck in summer 2012. All interested in participating should contact [Przemek Lenkiewicz](#) about it.



Upcoming Projects with TLA participation

by Peter Wittenburg

Participation in externally funded projects is very important for the TLA (The Language Archive) team for the usual reasons: (1) ensure funding to maintain existing software and add new functionalities – both being essential to maintain software; (2) participate in open competitions to show and to improve competence; (3) open new opportunities in a dynamic IT landscape. In this respect TLA was very successful during the last months, although the effort to form stable consortia and to come to proper proposals was considerable. We were part of 6 proposals from which 5 were accepted. It is a pity that the CLARICLE proposal which was meant to support the CLARIN ERIC in its construction efforts was not accepted.

CLARIN D (BMBF) Common Language and Technology Research Infrastructure 2011 – 2016

The follow-up project for the German D-SPIN (CLARIN) has been granted and will start officially at 1.5.2011. The new CLARIN D will participate in building the language resource and tools infrastructure and is therefore part of the European CLARIN ERIC initiative which will become a legal entity in 2011. In this initiative TLA will become one of the strong centers, improve some of the already started frameworks and add new ones that will turn out to be important for building and maintaining a useful research infrastructure enabling e-Humanities. Since we have reported frequently about CLARIN we refer for further information to the web-site.

DASISH (EC) Data Service Infrastructure for the Social Sciences and Humanities 2011 – 2014

This project brings together all 5 ESFRI research infrastructure initiatives in the social sciences and humanities (SSH) represented each by some centers: CLARIN, DARIAH, CESSDA, ESS, SHARE. The goal is to determine areas of possible synergies in the infrastructure development and to work on a few concrete joint activities. The rationale behind this idea is that a) double developments should be prevented, b) initiatives should mutually benefit from the advanced work of the others and c) to establish joint integrated domains where this makes sense for the SSH users. Joint activities will be along the following dimensions: understanding the different architectural solutions, assessing and improving data and metadata quality, setting up a tools and services forum, improve the quality of survey data, locate and improve data preservation and curation services, develop a joint shared data access and enrichment framework (AAI, PIDs, joint Metadata, Workflow implementations, joint annotation framework), jointly work on legal and ethical aspects, carry out much training and education work, work on disseminating the results. For TLA



this is a very interesting opportunity to disseminate resources and tools to other disciplines and integrate good components from others in the CLARIN infrastructure. This project is expected to start after the summer time in 2011.

INNET (EC) *Innovative Networking in Infrastructure for Endangered Languages* 2011 – 2014

This project will strengthen our international activities which were started in the DOBES project on the one hand and in CLARIN on the other. Together with the University of Cologne and colleagues from Poznan and Budapest we will start the following activities in the area of endangered language documentation and archiving: (1) setup 3 new regional archives and run annual workshops with all experts active in the current and coming regional centers; (2) organize best practice meetings with international guests and summer schools, (3) work out educational material to go into schools to get pupils' attention. In all infrastructure aspects the CLARIN agreements will be of relevance. For TLA it is an excellent opportunity to extend its archiving network and it is of course of great importance to spread the CLARIN messages. More about this project will be said in a separate article. This project is expected to start in June/July 2011.

EUDAT (EC) *European Data Infrastructure* 2011 – 2014

EUDAT is a first consequence of the report "Riding the Wave" of the EC's High Level Expert Group on Scientific Data in so far as it brings together 13 community driven infrastructure initiatives and 10 data centers to build a first prototype of a Collaborative Data Infrastructure (CDI). In such a CDI the community infrastructures take care of user oriented services on data, the data centers take care of common horizontal data services which are the same or at least very similar for all research disciplines and where both need to address topics such as data curation and establishment of trust between all stakeholders. CLARIN is one of the communities being selected in this project of strategic relevance. It has been understood worldwide that our efforts to take care of research data in terms of their preservation and in order to maintain accessibility need to be strengthened. Therefore EUDAT will focus on professional and robust common services such as: (1) providing an easy deposit for all involved researchers, (2) setup a distributed architecture allowing the participating centers to easily store large data volumes for preservation and access purposes (which includes a safe replication of data), (3) working on a policy-rules based replication at logical level of collections, (4) testing generic web services execution frameworks. This project is expected to start at 1.10.2011.

Radieschen (DFG) *Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur* 2011 – 2014

This project can be compared with the EUDAT project in so far as it tries to define the basis and roadmap for a future data infrastructure for the research



domain in Germany. Whiel EUDAT is already meant to come up with concrete services, Radieschen will make many interviews with experts from different stakeholders which will be analyzed in a few major dimensions with the goal to come up with a suggestion how the Collaborative Data Infrastructure can be realized in Germany with its federal organization structure. This project will start at 1.5.2011

References:

- More information on the existing Regional Archives set up by TLA can be found in [our flyer on the topic](#) (pdf).
- More information on CLARIN can be found on the [CLARIN web site](#).
- More information on the CLARIN ERIC can be found in [this issue of the CLARIN newsletter](#) (pdf).
- Additional information on PARADE the project idea that became EUDAT can be found at the [PARADE web site](#).
- More information on EUDAT can be found in [this presentation by Kimmo Koski](#) (ppt).
- Riding the Wave – How Europe can gain from the rising tide of scientific data” is the final report of the European Commission’s High Level Expert Group on Scientific Data. It is [available for download](#) (pdf).

[Introduction of a new transcription mode in ELAN 4.1.0](#)

by Aarthy Somasundaram & Han Sloetjes

In this new release of ELAN a completely new “Transcription Mode” and an improved “Segmentation Mode” are introduced. Both have been developed in close cooperation with [ELAN users](#). The Transcription Mode is built for high-speed transcription. Where the traditional Annotation mode can be seen as accuracy-oriented rather than productivity-oriented, the Transcription mode aims at increasing the speed and efficiency of transcription work. The user interface has been designed with convenient text entry in mind: the main element is a table containing the annotations of selected tier types, displayed in a vertical order. Each cell in the table represents an annotation (or a position where a depending annotation can be created). The segments (annotations) need to be created first, in the segmentation or annotation mode, after which text can be typed into the (empty) segments in this mode. Operation in this mode is very much keyboard oriented. Selecting an annotation plays the corresponding segment automatically and brings it into edit mode: ready for you to start typing. Press the TAB key to replay. After editing, hit ENTER (or use the navigation keys) to jump to the next annotation, to play that segment automatically and to start typing right away and so on... Activation of a cell will silently create child annotations if they don’t exist yet — merely clicking an

empty cell (or moving there using the keyboard) creates an annotation and opens it for editing. All this brings down the transcription work to just listening and typing, making it easy for the transcriber.

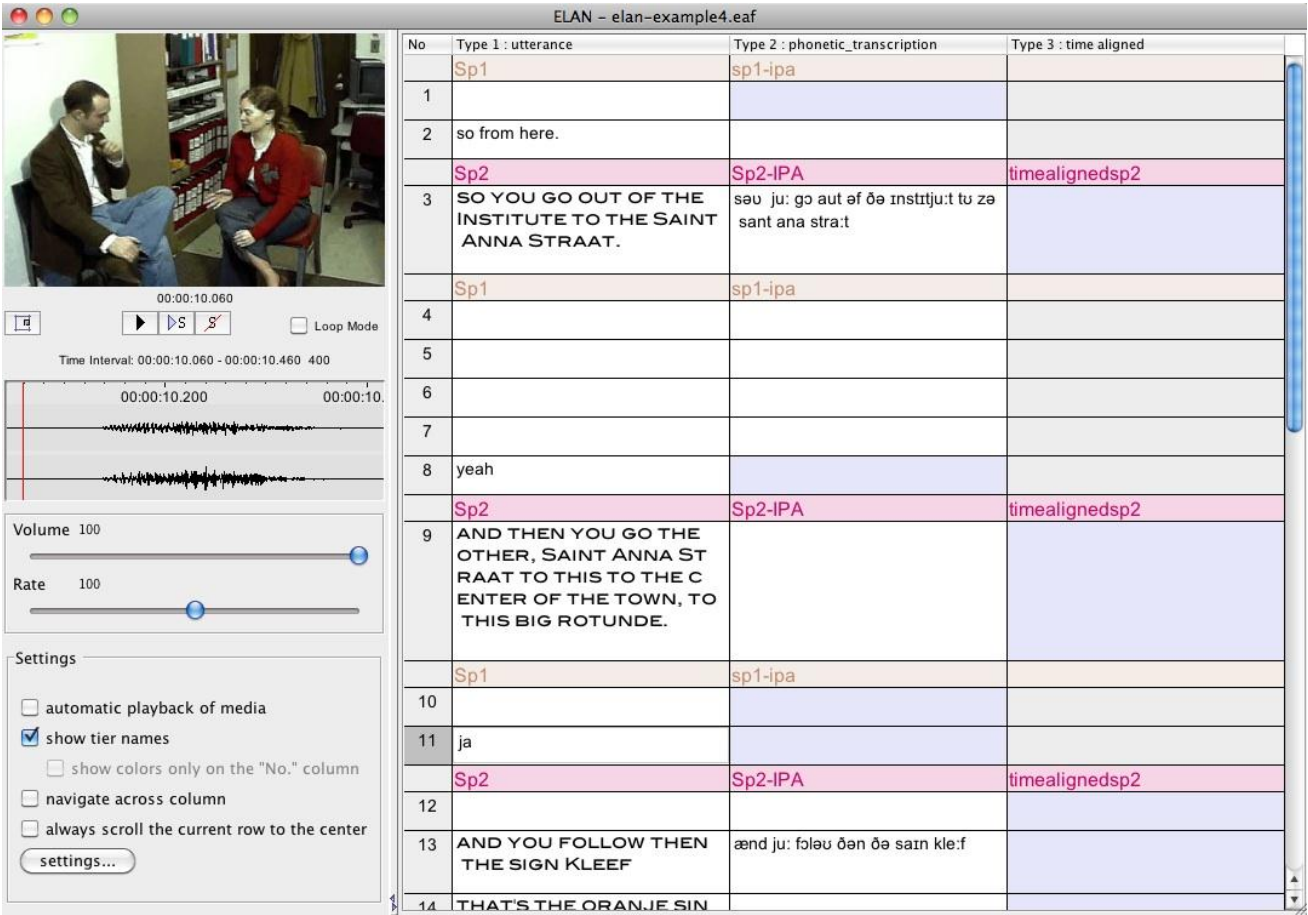


Figure 1: The Transcription Mode

“On-the-fly Segmentation” has been moved into the main window as the new Segmentation mode (instead of in a separate dialog). It is now easier to switch between tiers while the media is playing. Segments are created by keyboard strokes and can be modified by dragging with the mouse. This mode introduces a preliminary step-and-repeat playback mode.

Apart from that, some new multiple file processing functions have been added, like annotations from overlaps and annotation statistics. An option to add a group of tiers for a new participant has been implemented, as well as for deletion of multiple tiers in one action. Customization of the program has been improved by the introduction of new preference elements.

The new version can be downloaded at the [ELAN web site](#) where you will also find the updated [manual](#), detailing how to use the new modes and other new



functionalities.

Semantic interoperability of linguistic resources now and in the future

by Menzo Windhouwer

Language resources are a very valuable asset. Not only now, where they form the basis for new scientific publications, but also in the future when new research might need to reassess previous findings. Primary data, like audio and video recordings, can by the curation efforts of the archive managers still be accessible in this future. However, for a lexicon or a grammatical description curation is not so easy. The semantics of the terminology used by the creators of these resources can have drifted off, i.e., the terms might now have a (slightly) different meaning. So it is easily possible that future users have a hard time interpreting the resource in the right way or even come to wrong conclusions based on wrong assumptions. A possible solution would be to make the semantics associated with these resources explicit. The Data Category Registry, nicknamed ISOcat, is taking that route.

ISOcat provides a way for resource creators to describe and share the semantics of the elementary descriptors, called data categories, in their resources. Each data category becomes uniquely identifiable by a so called persistent identifier. And as the name of this identifier indicates, data categories in this registry are meant to stay around for a very long time. Future researchers should thus be able to take a resource from an archive and resolve these identifiers to get to the semantic descriptions of the data categories used in the resource. These descriptions should then help this researcher to interpret the resource.

However, already now adding data category identifiers to resources can help us. Because data categories can be reused by various resources they provide hints on which resources are semantically close together, i.e., they can help researchers to find more interesting resources based on semantic closeness. In these cases islands of resources using domain or application specific terminology can be connected as the specification allows the declaration of the use of various terms for the same data category.

ISOcat is the Data Category Registry for the [ISO Technical Committee 37](#), which develops many standards for linguistic resources. Standards like the [Lexical Markup Framework \(LMF; ISO 24613:2008\)](#) and the, in preparation, [Linguistic Annotation Framework \(LAF; ISO/DIS 24612\)](#) rely on the use of data categories taken from this registry to turn an abstract model into a model that is actually useful for a specific resource (type). The ISO committee is working towards sets of standardized data categories for various domains, e.g., metadata and



morphosyntax. This work is reflecting in ISOcat as public accessible Thematic Views. However, every linguist can actually create her own data categories, share them with others and offer them for standardization. This grass roots approach aims at providing a standardized core useful for a broad range of linguists, and reusable data categories for and maintained by specific groups of linguists.

Tools provided by [The Language Archive](#) are starting to interact with ISOcat. In [ELAN](#) items in a controlled vocabulary can be taken from ISOcat. [LEXUS](#), which allows the construction of LMF compliant lexica, can interact with ISOcat to select data categories to actually instantiate the abstract LMF data model. The [Component Registry](#) allows elementary elements and values in [component metadata](#) to link to ISOcat data categories. While these are just first steps and more will be needed the ultimate goal is that this will support the semantic interoperability of linguistic resources and thus research now and in the (far) future.