Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2644–2648
Marseille, 11–16 May 2020

Processing Language Resources of Less Resourced and Endangered Languages for the Generation of Augmentative Alternative Communication Boards

Anne Ferger
University of Hamburg
Germany
anne.ferger@uni-hamburg.de

Abstract

Under-resourced, under-studied and endangered or small languages yield problems for automatic processing and exploiting because of the small amount of available data as well as the missing or sparse description of the languages. This paper shows an approach using different annotations of enriched linguistic research data to create communication boards commonly used in Alternative Augmentative Communication (AAC). Using manually created lexical analysis and rich annotation (instead of high data quantity) allows for an automated creation of AAC communication boards. The example presented in this paper uses data of the indigenous language Dolgan (an endangered Turkic language of Northern Siberia) created in the project INEL(Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages) (Arkhipov and Däbritz, 2018) to generate a basic communication board with audio snippets to be used in e.g. hospital communication or for multilingual settings. The created boards can be imported into various AAC software. In addition, the usage of standard formats makes this approach applicable to various different use cases.

Keywords: language resources, endangered languages, augmentative alternative communication

1. Introduction

(Re-)Using cost-intensively generated language resources is a question of sustainability as well as economic efficiency. By exploiting the advantages of enriched language resources created in linguistic research projects while working with standard formats, a more wide-spread use of the language resources can be achieved.

The approach described in this paper uses a standard format of linguistic transcription data of an endangered language and generates standard format communication boards with audio snippets to be imported into various alternative augmentative communication (AAC) software. AAC includes various forms of communication for people with different limitations to their communication abilities. This includes low-tech (such as card-board communication boards or books with symbols for individuals to point at for communication) as well as high-tech (such as voice output communication aids (VOCAs) with eye-tracking functionality) solutions. The choice of AAC devices and their application is heavily dependent on the individual needs and abilities of the users. Simple communication is also often used in hospital settings where the ability to communicate can e.g. be temporarily obstructed by medical devices. Furthermore, simple communication boards can be used in multilingual settings where emergency basic communication is needed [1] .

The INEL Dolgan Corpus (Däbritz et al., 2019) used for this task is available under the licence CC BY-NC-SA 4.0 (public)[2]. Dolgan, a Turkic language spoken on the Taymyr peninsula and in adjacent areas in Northern Siberia has approximately 1,000 speakers.

While the example communication board created in this paper functions as a proof of concept and its real life usage may not be widespread, the general workflow can be used in various settings, especially for emergency (hospital) communication and AAC for small and under-resourced languages.

2. Related Work

While the processing of under-studied and under-resourced languages is becoming an important topic in computational linguistics (e.g. using natural language processing (NLP) techniques (Ren et al., 2014)), AAC for small languages is an area that still requires further work and research. AAC software (e.g. Coughdrop [3]) often uses various text-to-speech APIs which mostly exist for widely spoken and extensively resourced languages only. By taking advantage of the alignment of the transcription to audio snippets communication boards with speech output functionality can be created without the need for text-to-speech functionality.

Furthermore the boards mostly need to be created manually. Using language resources to generate communication boards automatically could decrease the manual work involved in creating communication boards.

Automatically generating AAC board using linguistic research data enhances the re-usability and sustainability of those resources and opens them for different domains and use cases. In the following related work concerning digital communication boards with audio output and the exploitation of language resources for endangered and under-resourced languages will be discussed.

2.1. Exploiting Language Resources of Endangered Languages

Exploiting language resources for various applications and software is facilitated by the use of standard formats and the

[1] e.g. in refugee camps https://www.tobiidynavox.com/support-training/downloads/boardmaker/refugee-communication-boards/

[2] PID:http://hdl.handle.net/11022/0000-0007-CAE7-1

[3] www.mycoughdrop.com

existence of only explicit information in the language resources. By using standards and standard formats, routines to use language resources for specific applications can be applied to other use cases and language resources as well. One standard for linguistic corpora is the TEI-format [4]. For spoken language there exists an additional ISO standard on how the TEI should be structured [5] (ISO/TC 37/SC 4, 2016).

Most processing functionalities for language resources (e.g. NLP techniques) work with big amounts of unordered and unanalyzed data, for e.g. automated translation functionality (e.g. (Bowker, 2000)).

While some work is done to use language resources for research and applications for people with disabilities (see e.g. (Yaneva et al., 2017)), combining endangered languages and research on speech pathology is not an obvious field of research. Nevertheless, the findings of research on endangered languages can have an important impact on the work of speech pathology, see (Ball and Bernhardt, 2008). (Small) language resources created for endangered and under-studied languages can be exploited for various applications for people with limitations, especially when utilizing further information and annotations in the resource instead of focusing on a large amount of data.

2.2. Creating Digital Communication Boards with Audio

In speech pathology the term AAC (Augmentative Alternative Communication) is used for various forms of communication for people with disabilities or other restrictions affecting their ability to communicate. Communication boards represent a low tech solution for AAC. Those boards can consist of analogue paper boards with symbols or images and the respective words or lexemes written next to them or of a digital version with audio output functionality and also a multi-page layout accessible through folders. Basic communication boards belong to low-tech AAC solutions to assist people with special communication needs interacting with other people.

The use of the boards is e.g. common in hospital settings and critical care, see (Patak et al., 2006) as well as in e.g. refugee facilities [6]. Furthermore boards can be used for aided language stimulation (Jonsson et al., 2014).

While many resources for facilitating the creation of communication boards exist [7], the automatic generation of the boards is not common. One reason for that is the high amount of individual adaption that is needed for many use cases of the boards, depending on the needs and abilities of the users.

While there are use cases where individual changes to the boards are not needed or not possible (esp. in emergency communication) the automatically generated boards still allow for manual adaption to individual needs afterwards when an adaptable standard format is used.

As established before, using standard formats is crucial in modern research. Since a lot of AAC technology is available in proprietary formats only, the Open AAC initiative [8] aims to make the use of open-source standard formats more wide-spread. For communication boards the Open Board Format (.obf) [9] is the proposed format. The format can be imported in and exported from various AAC software already[10]. The generation of the boards happens manually in one of the applications, either by adapting existing available free boards [11] or creating completely new boards, which demands a high amount of manual work.

The Open Board Format consists of JSON files which describe the structure and layout of a communication board, composed of buttons, images and a grid-based layout for the buttons [12]. While most apps use text-to-speech functionality to generate audio output, the open board format also allows for audio snippets to be stored with certain buttons indicating words or lexemes.

Some basic communication boards do not deal with inflection, which still produces understandable output for many languages, depending on their structure and typology, while others include either manual or semi-automated inflection functionality. The app Coughdrop uses a sidebar with static inflection indicators while e.g. the proprietary software Tobii Sono Lexis[13] automatically inflects verbs depending on the previously chosen lexemes. While it is important to adapt the boards to individual communication needs for every user, the automatic generation, especially for smaller languages, can help to decrease the amount of manual work related to the creation of individually adapted boards.

This approach shows a proof of concept to automatically insert information from a language resource using a template of a digital communication board containing very basic lexemes that can be used for basic communication. For this template an openly available board from the Project Core [14] was used containing 36 symbols and words usable for very basic communication[15]. Of course the nature of the structure and typology of a language plays a big role in the way such a template should or can look like, but for the sake of simplicity the following will just focus on an English template board and its adaption to the Dolgan language resource.

[4] https://tei-c.org/

[5] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338

[6] see e.g. Icoon for refugees (http://icoonforrefugees.com/) for an approach of using visual icons for communication in refugee settings

[7] see e.g.
CoughDrop (https://www.mycoughdrop.com/)
Cboard (https://www.cboard.io/)
The Open Voice Factory
(http://www.theopenvoicefactory.org/)
Picto4me (http://www.picto4.me/)

[8] https://www.openaac.org/

[9] https://www.openboardformat.org/

[10] see footnote 7

[11] https://www.openboardformat.org/examples

[12] see https://www.openboardformat.org/docs

[13] http://www.rehamedia.de/fileadmin/downloads/Handbuecher/Tobii/Tobii_Sono_Lexis_Manual_German.pdf

[14] project-core.com

[15] https://app.mycoughdrop.com/wahlquist/projectcore-36universalcore

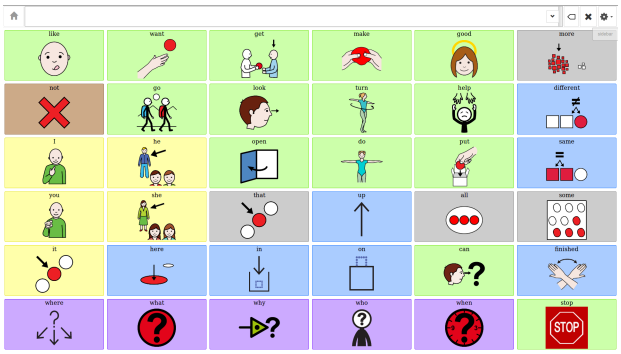

Figure 1: Example of a communication board: Project Core 36 universal core.

3. Generating Communication Boards Using Language Resources of Under-Resourced Languages

Automatically generating digital communication boards in standard format with audio snippets solves the problem of missing text-to-speech functionality for small or under-resourced languages. The format of the available Dolgan corpus is TEI[16] following the ISO Standard for spoken language (ISO/TC 37/SC 4, 2016) with the corresponding audio files referenced via time point anchors. Using English glosses that are additionally present in the resource, an English-based template communication board could be used for the generation.

For the template board an existing available board in the Open Board format [17] was used. By exploiting the glossing and part of speech information available in the Dolgan data the linking between the English template and the Dolgan data could be established.

Using XQuery and XSLT technology those enriched files were searched and specific words or lexemes including their respective audio snippets were added to the template board. The generated board in the OBF format could then be imported into AAC software such as Coughdrop and used for very basic communication in the Dolgan language.

3.1. Advantages of the Approach

Exploiting a resource of an endangered language to create a resource to be used in the AAC context can open the resource to wider scopes of application and simultaneously decrease the amount of manual work needed for the creation of digital AAC communication boards. Furthermore the problem of missing text-to-speech API for small or endangered languages that would be needed for such boards can be solved by automatically extracting audio snippets from the language.

Using standard formats for this workflow allows for various other applications as well. In the field of endangered languages, some use cases could also be the creation of material to help revitalization of languages or for teaching purposes.

[16] https://tei-c.org/

[17] https://www.openboardformat.org/

3.2. Requirements of the Language Resource

The INEL Dolgan Corpus (Däbritz et al., 2019) used for this example is available in the TEI format following the ISO standard for spoken language (ISO/TC 37/SC 4, 2016). The language resource contains Dolgan transcriptions with various annotations and translations (e.g. into English). The annotations needed for the described workflow are underlying morphophonemes, English glossing and part of speech tagging. The transcriptions in the resource are aligned to audio recordings which is a requirement for the audio functionality of the generated board.

In the resource described in this paper the transcriptions are aligned to the respective audio files, the underlying lexical forms of morphemes are annotated in the tier 'mp' and the English glosses are annotated in the tier 'ge' and the part of speech information was tagged in the 'ps' tier[18]. The various annotations relate to different parts of the transcription (either morphemes, words or sentences) which will be discussed further in the technical details of the workflow.

3.3. Template for a Communication Board

For a template an openly available board from the Project Core [19] available in the Open Board Format was used. It contains 36 symbols and words usable for very basic communication[20]. The approach behind the creation of the board follows the core vocabulary theory (see (Beukelman et al., 1989) and (Banajee et al., 2003)). The board containing 36 vocabulary items was chosen for the sake of simplicity while still enabling real life use cases.

The chosen board should serve as a template in a twofold way: Firstly to serve as an actual technical template for the extracted language resource to be inserted, secondly to serve as a more abstract template to be substituted with various different boards in the obf format.

3.4. Automatic Generation of the Communication Board Using the Language Resource

The workflow to generate a Dolgan digital communication board was to create an XSLT transformation on the template JSON file to fill it with information from the TEI XML files following the ISO standard for spoken language. To achieve this, the English glosses in the template board (which also have part of speech information encoded in the format) were matched with elements in the 'ge' tier of the TEI file, then the part of speech annotations in the 'ps' are matched with the part of speech information stored in the template board (to ensure the correct matching of the words or lexemes). When a correct match is found the value of that match in the 'mp' is used to ensure the selection of the correct base form. Because of the alignment to the audio in the transcription files the audio snippets can be cut

[18] for further documentation and information of the annotations present in the Dolgan corpus see https://corpora.uni-hamburg.de/hzsk/de/islandora/object/file:dolgan-1.0_INEL_Dolgan_Corpus_1.0_User_Documentation/datastream/PDF/INEL_Dolgan_Corpus.pdf

[19] project-core.com

[20] https://app.mycoughdrop.com/wahlquist/projectcore-36universalcore

automatically and stored for the integration in the communication board. The Open Board Format additionally offers a zipped obz format that can contain the audio snippets as files. After the generation the file can be validated using the Open Board Format Validator [21] and imported into one of the compatible AAC applications.

3.5. Technical Details

The generation was carried out by an XSLT transformation using the JSON template and adding the information from the resource using XPath and XQuery functionality. The stylesheet will be made available as open source in the future.
The found locations in the transcriptions can then be used to extract the audio that should accompany the words on the communication board. The audio snippets can be linked in the zipped Open Board Format (.obz) by providing the path to the respective snippet in the images folder.
The search for one lexeme uses the label from the English template board to search the same entry in the English Glossing tier ('ge') of the language resource (on morpheme level). When a matching instance is found it compares the part of speech information in the English template with the part of speech annotation ('ps' tier) of the word (that is linked with the morpheme that was found) in the language resource. If this information also matches, the value of the morpheme in the annotation tier 'mp' (underlying lexical forms of morphemes) is chosen as a possible value to be inserted in the template board. If varying values are found the value occurring most frequently is chosen.

```
let $tier1:="ps",$tier2:="ge",
$tier3:="mp",
$lemma:="can",$ps:="v",
$mid:=//spanGrp[@type=$tier2]/span/
span[text()=$lemma]/@from,
$wid:=$mid/../../@from
return //spanGrp[@type=$tier1]/
span[@from=$wid][text()=$ps]/../
../spanGrp[@type=$tier3]/span/
span[@from=$mid]/text()
```

Figure 2: XQuery used to find instance of "can" with the part of speech tag verb (v)

The produced obf file was then validated using the File Validator[22] from the Open Board Format website and converted into a pdf file using the Preview Generator[23] from the Open Board Format website. It could also be imported into various other tools.

3.6. Example of a generated Dolgan Board

The following 3 and 4 show the pdf preview version of the generated board with the Dolgan language input. The same obf file imported into e.g. the CoughDrop software allows for the choice of a sequence of different buttons and the audio output of the chosen buttons.

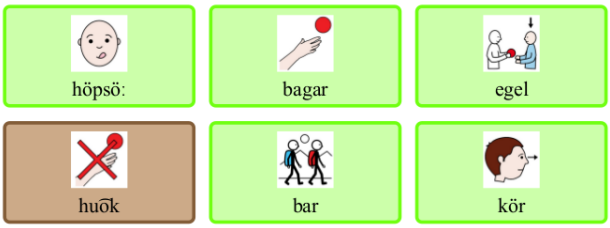

Figure 3: Extract of the generated Dolgan communication board.

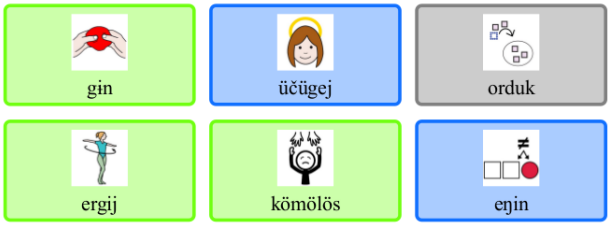

Figure 4: Extract of the generated Dolgan communication board.

3.7. Encountered Difficulties

While the pipeline of the processing of the language resource and the JSON seemed technically straight-forward in the beginning, many obstacles were found during the generation of the example board.
On the one hand high quality resources are needed that contain (English) glosses, part of speech tagging as well as underlying morphophonemes forms. On the other hand, the resource needs to be large enough for the wanted forms to occur in the data.
The nature of glossing and its inexactness also call for further lexical analysis in the data to ensure a high quality of the generated output files. While inflections don't play a role in this example, they should in future developments of the approach. Both of these difficulties could be overcome by using an additionally created lexical resource for the language resource.
Furthermore the alignment of the transcriptions and the audio files needs to be very exact and thorough, especially on the level of morphemes, which is usually hard to achieve. While it is possible to automatically extract the correct snippets of audio for the lexemes the result can not compare to the usage of a text to speech API because the snippets come from different contexts, speakers, are available in different quality and don't blend together in a sentence.
The last difficulty encountered while testing the approach were real life use cases. While the example created was not meant to be more than an example, the template for the communication board needs to be adapted to specific use cases in the future.

[21] https://www.openboardformat.org/tools
[22] https://www.openboardformat.org/tools
[23] https://www.openboardformat.org/tools

4. Outlook

While the proposed approach in this paper showed some difficulties in its application it also showed a lot of potential to adapt the workflow and the resources involved to yield better results. Using lexical resources additionally to the language resources on hand will be possible with further developments in the INEL project. The high quality in the resources needed for the workflow is a fundamental requirement for scientific language resources and not only of the described approach.

While the use cases are limited because of the nature of the template board, the adaption of the template board is a necessary step to allow for specific use cases, either emergency communication in small languages or facilitating of community involvement for endangered languages. For specific future use cases the exploitation of metadata linked to the resource also has great potential for e.g. context-specific boards.

Another approach to enhance the workflow could also be the application of NLP technology for small languages.

Possible further use cases could be the generation of material to help revitalize or teach endangered languages as well as providing basic help for communication for people that do not speak the language, e.g. researchers or travelers.

5. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

6. Bibliographical References

Arkhipov, A. and Däbritz, C. L. (2018). Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology*, (3):9–18. http://ling.tspu.edu.ru/en/archive.html?year=2018&issue=3&article_id=7130.

Ball, J. and Bernhardt, B. M. (2008). First nations english dialects in canada: Implications for speech-language pathology. *Clinical Linguistics & Phonetics*, 22(8):570–588.

Banajee, M., Dicarlo, C., and BURAS STRICKLIN, S. (2003). Core vocabulary determination for toddlers. *Augmentative and Alternative Communication*, 19(2):67–73.

Beukelman, D., Jones, R., and Rowan, M. (1989). Frequency of word usage by nondisabled peers in integrated preschool classrooms. *Augmentative and Alternative Communication*, 5(4):243–248.

Bowker, L. (2000). Towards a methodology for exploiting specialized target language corpora as translation resources. *International Journal of Corpus Linguistics*, 5(1):17–52.

ISO/TC 37/SC 4. (2016). Language resource management – Transcription of spoken language. Standard ISO 2462:2016, International Organization for Standardization, Geneva, CH.

Jonsson, A., Kristoffersson, L., Ferm, U., and Thunberg, G. (2014). The ComAlong Communication Boards: Parents' Use and Experiences of Aided Language Stimulation.

Patak, L., Gawlinski, A., Fung, N. I., Doering, L., Berg, J., and Henneman, E. A. (2006). Communication boards in critical care: patients' views. *Appl Nurs Res*, 19(4):182–90.

Ren, Y., Kaji, N., Yoshinaga, N., and Kitsuregawa, M. (2014). Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. *IEICE TRANSACTIONS on Information and Systems*, 97(4):790–797.

Yaneva, V., Orăsan, C., Evans, R., and Rohanian, O. (2017). Combining multiple corpora for readability assessment for people with cognitive disabilities. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 121–132, Copenhagen, Denmark, September. Association for Computational Linguistics.

7. Language Resource References

Däbritz, Chris Lasse and Kudryakova, Nina and Stapert, Eugénie. (2019). *INEL Dolgan Corpus*. Archived in Hamburger Zentrum für Sprachkorpora. Wagner-Nagy, Beáta; Arkhipov, Alexandre; Ferger, Anne; Jettka, Daniel; Lehmberg, Timm, The INEL corpora of indigenous Northern Eurasian languages, 1.0 Publication date 2019-08-31. http://hdl.handle.net/11022/0000-0007-CAE7-1.