An Investigation of the Grammar Matrix, AGGREGATION Project, and LING 567 in the Context of Value Sensitive Design

Tara Wueger

Department of Lingusitics, University of Washington
Box 354340, Seattle, WA, 98195-4340
taraw28@uw.edu

Abstract: In this paper, I explore the possible impacts of the Grammar Matrix and AGGREGATION project within the contexts of how they are used in the University of Washington's grammar engineering class (LING 567) and how they could be used outside of the classroom, as well as taking into account datasets that make up the AGGREGATION data repository. I use methods from value sensitive design, doing a stakeholder analysis (on both direct and indirect stakeholders) and detailing value scenarios that highlight societal implications of technology using Matrix-produced grammars. I conclude that stakeholders, like speakers and linguists, should be more involved in the grammar engineering process when they want to be and that the AGGREGATION data needs to be better documented in order for the interests of language communities to be better represented.

Keywords: grammar engineering | machine translation | value sensitive design | local languages | educational context

1 Introduction

The LinGO Grammar Matrix customization system (Bender, Flickinger & Oepen 2002, Bender et al. 2010, Zamaraeva et al. 2022) allows someone to build a grammar for a language by answering questions and making selections based on the language. It is one part of the AGGREGATION project (Bender et al. 2013, 2014, Howell et al. 2017, Howell & Bender 2022, Zamaraeva et al. 2017, Zamaraeva, Howell & Bender 2019), which involves using data to automatically create a machine-readable grammar for a language and using said grammar to parse and translate sentences. The Matrix is one of the main tools used in LING 567, which students use in order to incrementally build up a grammar for a specific language. In this paper, I investigate the following research questions:

1. Could LING 567's approach to grammar engineering be improved?
2. What are the possible impacts of involving speakers, communities,[1] and linguists in the grammar engineering process? Are there impacts of not involving them?

[1] I distinguish between speakers and communities, where speakers are individuals who speak a language and communities are groups of these speakers, because sometimes it is important to distinguish what an individual might think from the general attitudes of a community.

First, I identify the direct and indirect stakeholders for Matrix grammars and then identify value scenarios that highlight the above-stated questions. I find that stakeholders include speakers, communities, linguists, students, and grammar engineers and that these stakeholders can be affected differently depending on how widely spoken a language is. Additionally, I find that it is important that grammar engineers work with the communities of the languages they build grammars for so that their culture can be properly represented. It is also important for those from other fields or disciplines, like Natural Language Processing (NLP), to work with speakers, communities, and linguists so that the wishes of the speakers and communities are respected.

2 Positionality Statement

To give context for my interest in the topic of this paper, I took LING 567 as a student in 2022, have done work in the AGGREGATION project, and am currently working on a thesis that is part of the Grammar Matrix. As my education has progressed, I have become increasingly interested in the ethics of NLP systems, which is what has led to the creation of this paper. One important thing to address is that, as a student and white woman whose native languages are English and German (two highly-studied standardized languages), I will have a fairly detached view on the subject matter. I am not a member of any of the communities for which we have data for and am not able to interact with the speakers/communities/linguists themselves. Additionally, I do not have the experience of speaking a language whose community has a history of being marginalized. The goal of this paper is to start the conversation about speaker-, community-, and linguist-involvement in the AGGREGATION project, Matrix customization system, and LING 567 class. I do my best to consider a variety of sources and viewpoints, however, I recognize that there will be things I don't consider due to my background.

3 Background/Related Work

3.1 The AGGREGATION Project

The Automatic Generation of Grammars for Endangered Languages from Glosses and Typological Information (AGGREGATION) project is a system that can create grammar specifications from Interlinear Glossed Text (IGT) corpora. The pipeline for this system can be seen in Figure 1. The format of the data at the initial stage of the pipeline is IGT, which is the standard format used by linguists for example text (usually sentences or phrases). This is then converted to XIGT (Goodman et al. 2015), an XML-based format for representing IGT. INTENT (Georgi 2016) is used to enrich the XIGT so that it has part of speech tags and syntactic dependencies. Next, the BASIL inference system, using the Matrix-Odin Morphology (MOM) system (Wax 2014, Zamaraeva, Howell & Bender 2019), outputs a grammar specification file (or choices file). This choices file contains options that correspond to information about the grammar of a language (e.g. what morphological affixes a language uses). This file can also be fed into the Matrix webpage, which will be pre-populated with the choices from the file and allow a grammar engineer to make any changes to the choices generated by BASIL or make additional decisions about phenomena not handled by BASIL. Alternatively, a grammar engineer can start from scratch and manually fill out

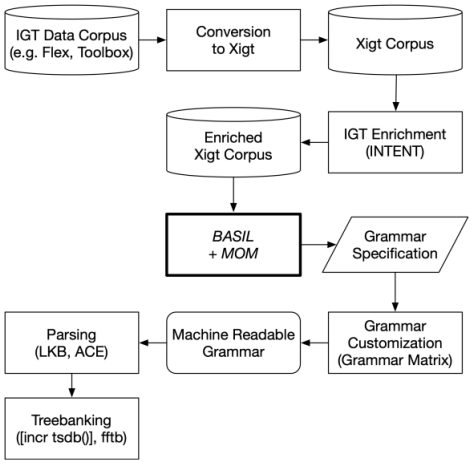


Figure 1: AGGREGATION pipeline (Howell & Bender 2022: p. 4)

the Matrix questionnaire. Once the questionnaire has been filled out (to the extent that is needed[2]), the Matrix customization system can produce a machine-readable Head-Driven Phrase Structure (HPSG) grammar (Pollard & Sag 1994, Copestake 2002). Then, programs like the Linguistic Knowledge Builder (LKB) (Copestake 2002) and Answer Constraint Engine (ACE) (Crysmann & Packard 2012) can use the Matrix grammar to parse sentences, in order to determine whether they are grammatical and to see if the parses are correct syntactic and semantic representations of the sentences, as well as to generate sentences. Finally, the `[incr tsdb()]` (Oepen 2001) and Full Forest Treebanker (Packard 2015) systems can be used to further verify whether sentences have correct parses and find sources of ambiguity.

Although some members of the AGGREGATION project who have contributed data for a language have also recorded information for how the data can be used, however, others have not. The members that have provided this information are "outsider-linguists", meaning they are not members of the community that speak the language they are working on. It would be important to get into contact with the speakers, communities, or linguists to make sure the data is used appropriately and for resources they (meaning the speakers and communities) approve of as well as update the corresponding information.

3.2 The Grammar Matrix

By filling out the various subpages that make up the Grammar Matrix questionnaire[3] grammar engineers can generate a machine-readable grammar for a language of their choice. To get started, the grammar engineer needs to either be intimately familiar with the grammar of the language they are working with or have access to a resource grammar for that language. Then, they can go through each subpage (the full list of subpages can be found in Figure 2) of the questionnaire and fill it out based on their language. This process is an iterative one, as there are subpages that interact with one another, so the grammar engineer will most likely find themselves going

[2]It is not required for the entire questionnaire to be completely filled out. Beyond basic requirements (e.g. the lexicon), only phenomena that the grammar engineer wants to model need to be filled out.

[3]Hosted on `https://matrix.ling.washington.edu/customize/matrix.cgi`.

Figure 2: List of LinGO Grammar Matrix subpages

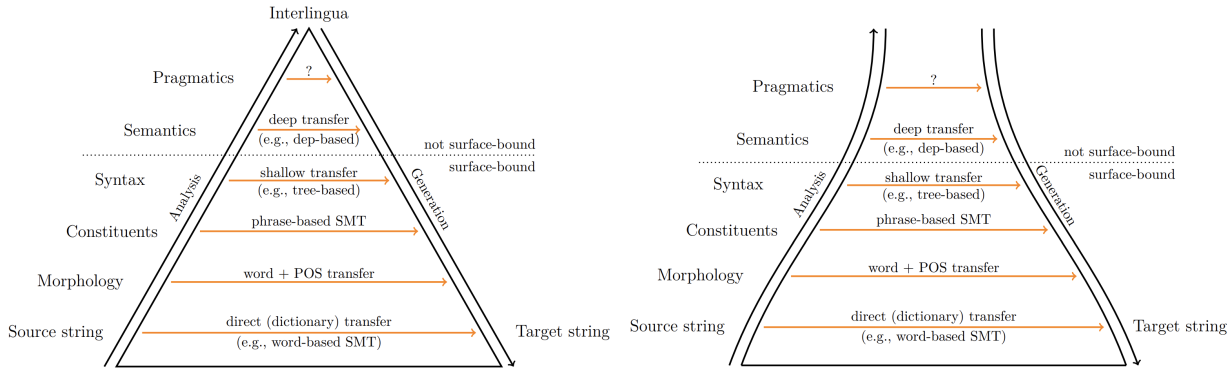
back and forth between them. Typically, it is easiest to start with the General Information and Lexicon subpages, followed by the Word Order, Number, Person, Gender, Case, and Tense, Aspect, and Mood subpages, and then moving on to the Morphology subpage. After filling out these subpages, they can implement the more complex phenomena covered by the other subpages. It is not necessary to fill out every subpage, as not all phenomena apply to every language, and because a grammar does not need every phenomenon in the language implemented.

When it comes to filling out a subpage, the grammar engineer is usually presented with different options. It is not uncommon to have multiple ways of implementing a given phenomenon, so the grammar engineer can choose the one they feel is most applicable based on their own knowledge and the analyses presented in the resource grammar they are using. Once they have finished filling out the subpages they want, they can download the grammar which contains many files, most of which have the `.tdl` file extension. These files use Type Description Language (TDL) syntax (Krieger & Schafer 1994). The grammar can then be used with software like the LKB, ACE, or `[incr tsdb()]` for parsing and generation (described above in Section 3.1).

3.3 LING 567

LING 567 - Knowledge Engineering in NLP - is a grammar engineering class currently taught in the Linguistics department at the University of Washington in Seattle, WA by Professor Emily M. Bender.[4] The class focuses on building linguistic grammars by hand using a variety of software (primarily the LinGO Grammar Matrix, the LKB, ACE, and `[incr tsdb()]`, discussed above

[4]See `https://courses.washington.edu/ling567/` for course information.



(a) Vauquois triangle (Goodman 2018: p. 7)

(b) Copestake volcano (Goodman 2018: p. 8)

Figure 3: Visual representations of transfer-based MT

in Sections 3.1 and 3.2), relying heavily on grammatical theory, specifically HPSG and Minimal Recursion Semantics (MRS), with an end-goal of a machine translation (MT) task. Part of the motivation behind this class is that these "precision" grammars tend to be more accurate than traditional machine learning (ML) methods, but are less robust (Bond et al. 2011: p. 88). However, incorporation of these grammars into ML methods has led to improved results, for example using precision grammars as training data for NLP systems (Bender & Emerson 2021: pp. 1125–1128) or using precision grammars in NLP applications for annotation tasks (Bender et al. 2015). Throughout the duration of the class, students build skills in incremental development of analyses, working with unfamiliar languages, testing and debugging linguistic analyses, and contributing to efforts in language documentation.

When it comes to using Matrix grammars in MT, LING 567 uses transfer-based MT. In order to translate from a source string to a target string using this method, an interlingua is often used - i.e. an intermediate language that the source string can be translated into and the target string can be translated from. This allows for scalability when creating an MT system between an increasing number of languages because a translation system for each language pair does not need to be designed. Instead, strings from each language need to be converted into the interlingua and strings from the interlingua need to be converted into each language. Figure 3a is a reproduction of what is commonly known as the Vauquois Triangle (Vauquois 1968), which is a visual representation of transfer-based MT. LING 567 uses a variation of this method where there is no intermediate language, instead using semantic representations in minimal recursion semantics (MRS). This is shown in Figure 3b, which is very similar to the Vauquois Triangle, except that there is no convergence to an interlingua, so it is represented with an open top – the "Vauqois inverted funnel with a very long spout" (Copestake 2007) or "Copestake volcano" (Goodman 2018: p. 8).

As described in Section 3.2, grammar engineers can use the Matrix to output a grammar that can be used (with the LKB or ACE) to parse sentences. The grammar engineer can also manually edit the files, known as "tdl-editing", which is needed when the implementation produced by the Matrix results in incorrect or ambiguous parses and when the Matrix does not contain a library for a phenomenon the grammar engineer wants to implement. One of the files contained in the grammar directory that is not a TDL file is the "choices" file. This plaintext file contains information extracted from the Matrix subpages that the grammar engineer filled out. If grammar engineers want

to make changes to the grammar, for example implement a new phenomenon, they can upload the corresponding choices file to the Matrix website, make changes there, and download a new grammar. If any tdl-editing was done to the original grammar files, the grammar engineer will need to compare the original files with the new files, copying the changes over manually. This is part of the process of improving a Matrix grammar, which focuses on coverage (the number of sentences that are parsed), accuracy (the number of sentences that are parsed correctly), and robustness (the number of parses per sentence). In this process, the grammar engineer will need to go back and forth between customizing with the Matrix and testing how the grammar performs on a testsuite (a collection of test sentences) in order to find and improve sources of ambiguity (resulting in a more robust grammar), incorrect parses (resulting in higher accuracy), and lack of parses (resulting in better coverage).

In addition to customization with the Matrix and tdl-editing, students work with the variable property mapping (VPM)[5] facility and create transfer rules, both of which are important for the machine translation part of the class. Since much of the information in grammars is encoded in variable properties and the grammar-internal and grammar-external variable properties can be different, the VPM is used to map from one to the other (in either direction). Transfer rules assist in the translation process and fill in the gaps left by using MRS as an interlingua by directly mapping some information from the source grammar to the target grammar.

Currently, the process of grammar engineering in LING 567 is very far removed from speakers, communities, and the linguists who gathered or analyzed the data. Students often have to make choices while implementing, and the thought processes behind these choices are not always linguistically motivated, instead being motivated by ease of implementation. Ideally, especially when the grammars are used beyond the classroom, speakers and communities should have input into such decisions or, when that is not possible, the linguists who worked with them should be consulted. Additionally, a choice is made when English is chosen as the glossing language in these projects. There are impacts of this choice and how it supports the idea that English is the "default" language.

3.4 Value Sensitive Design

Stakeholder analysis is one of the fourteen value sensitive design (VSD) methods detailed by Friedman, Hendry & Borning (2017). The purpose of this method is to identify individuals, groups, institutions, and societies that the technology being explored could affect, either positively or negatively. It involves identifying direct stakeholders (those interacting with the technology directly) and indirect ones (those being affected by the technology but not directly interacting with it) (Friedman, Hendry & Borning 2017).

Detailing value scenarios is another VSD method, first introduced by Nathan, Klasnja & Friedman (2007). The main goals of value scenarios are to come up with provocative scenarios that point out theoretical dangers of a technology. Listed below are the six concerns that should be considered when building new technology and can be illuminated with one or more value scenarios. They were identified by Friedman, Hendry & Borning (2017: p. 80), which build upon the five key elements identified by Nathan, Klasnja & Friedman (2007: pp. 2587–2588) (also below in parentheses):

[5]VPM is credited to Stephen Oepen but there is no official citation. Michael Goodman discusses VPM in his dissertation (Goodman 2018) and would be the best place to find more information.

a) implications for direct and indirect stakeholders (stakeholders)

b) key values (value implications)

c) widespread use (pervasiveness)

d) indirect impacts

e) longer-term use (time)

f) systemic effects (systemic effects)

3.5 Respecting Languages, Communities, and Speakers

One of the main themes of this paper revolves around outsiders of a community for a language they are researching respecting what members of those communities want, in terms of data use and technology produced from said data (which Ramponi 2024 coined as the "speaker-centric" approach). Speakers of what have commonly been referred to as "low-resource" languages have a history of being exploited, in more ways than one (see Taiuru 2023, Smith 2021, Dobrin, Austin & Nathan 2014). The concept of referring to languages as "high-resource" and "low-resource" is inherently problematic in that it can ascribe the worthiness of a language to the amount of machine-readable resources it has (Ramponi 2024). There has also been discussion in recent years involving problems with using death metaphors (i.e. endangered or extinct) to describe languages (Fostar 2021). Wes Leonard discusses something similar when describing issues with current classifications focusing on "(perceived) patterns of current use and not around the potential of future use" (Leonard 2008: p. 26). Bird (2022b) proposes a multipolar classification model – *standardized languages*, *local languages*, and *contact languages* – while recognizing that languages might belong to multiple of these categories and that there is room for improvement in this model. Within this model, *standardized languages* include "high-resource" languages and "low-resource" languages where commercial, social, or political resources are being used to create language technology, and the community wants this technology (Bird 2022b: p. 7817). On the other hand, *local languages* include small languages (usually indigenous or endangered) that are mostly oral, which are often called "low-resource" since there isn't a lot of existing data that can be used to create language technology (Bird 2022b: p. 7817). Lastly, *contact languages* include what are commonly referred to as trade and vehicular languages as well as languages that are used to communicate within a wider community, whose members often belong to various linguistic regions (Bird 2022b: p. 7820).

Linguists and those building language technology have had this ideal that indigenous languages need to be saved or revived (Dobrin, Austin & Nathan 2014), without taking into consideration if this is what the indigenous communities want or, if they do want it, how they want it done. A movement towards decolonization of speech and language technology is being spurred by speakers and communities. Smith (2021) does something similar in the second half of her book *Decolonizing Methodologies: Research and Indigenous Peoples*, drawing partially on her own experience as a Māori woman to suggest approaches for how researchers can work with indigenous communities. In Bird's (2020) "Decolonizing Speech and Language Technology," he identifies methods that linguists and technologists can use to build bridges with communities, starting with asking the communities to identify their goals. The aim of doing this is to create a world that sustains its languages, and not one that decreases language variability (Bird 2022a). People that work with

iso	Language	# of Speakers	Permissions
bcj	Bardi	<30	AGG, 567, Repro
ctn	Chintang	<5000	AGG (data-use agreement must be signed)
mni	Meitei	∼1,760,000	no data-use restrictions provided
yak	Yakima Sahaptin	25	no data-use restrictions provided

Table 1: Languages used in the stakeholder analysis
(all speaker counts come from Eberhard, Simons & Fennig (2023))

various languages can enrich the analyses or technologies they create by interacting with the communities that speak those languages. Again, this relies on respecting the people that we interact with and interacting with people from a diverse range of cultures and communities.

There are issues surrounding technology accessibility in indigenous communities. Toth, Smith & Giroux (2018) propose links between access to technology and problems with education and healthcare in indigenous communities. The base issue here is with limited or no access to computers or the Internet, not even addressing access to specific programs or software. Some of the main reasons behind the lack of access to technology in general is cost, computer literacy issues, and being unaware of how technology can be useful for indigenous goals or interests (which stems in part from some indigenous communities being geographically isolated from other communities) (Dyson, Hendriks & Grant 2006).

4 Methodology

In the following sections of this paper, I conduct a stakeholder analysis and detail value scenarios for the AGGREGATION project, the Grammar Matrix, and LING 567. There are 36 languages represented in the AGGREGATION data repository. For the stakeholder analysis, I select four languages that show a variety in number of speakers and data-use permissions in order to determine the direct and indirect stakeholders of a Matrix grammar for each of these languages (the summary of which can be found in Table 1). For value scenarios, I detail three scenarios: one that focuses on the possible negative effects of Matrix grammars on speakers and communities, one that shows negative interaction with the NLP community and potential societal impacts of such interaction, and one that shows the possible positive impacts of having two LING 567 students be in contact with the linguist who provided the data for the language they're working with and making implementation decisions based on feedback from that linguist. In the creation of these scenarios and the subsequent analyses of them, I keep the six concerns listed in Section 3.4 in mind and use them to direct the types of analyses I make about the scenarios and what they mean in a larger context.

5 Stakeholder Analysis

The stakeholders for all four languages were similarly identified. Students,[6] grammar engineers, and linguists were identified as direct stakeholders, since they are the ones that would directly interact with a Matrix grammar. Speakers, communities, and linguists were identified as indirect

[6]Although students are technically grammar engineers, I distinguish them from professional grammar engineers because stakes in the classroom differ from those in industry and academia.

stakeholders, since they could be affected by the technology without being directly involved. Linguists are in both groups because their stakes depend on their role. There are some that would directly work on a Matrix grammar (usually these linguists are also grammar engineers). Additionally, there are those that provided the data used for a Matrix grammar, created the resource grammar that is used when building a Matrix grammar, or used a Matrix grammar that had been created by a grammar engineer based on either the data or resource grammar they provided.

Another possible direct stakeholder would be a speaker or linguist from a community who is interested in this kind of work. The students in the LING 567 class (and the AGGREGATION project in general) deal mostly with local languages (see subsection 3.5), so if somebody that speaks such a language or a linguist who works with these speakers is interested in what the Matrix or AGGREGATION project can do for them, they could reach out and assist with providing data as well as work with developers. The other side of this is that a speaker or linguist could discover that their data is being used in ways they don't agree with. In this case, they could reach out to the developers and become active in the development process, hopefully coming to an agreement that everybody is happy with. If such an agreement is not possible, they could disallow all or certain uses of their data.

Analysis What differed between some of the languages is the kinds of technology that could be created and how it might affect stakeholders. For example, Meitei is a standardized language spoken by ∼1,760,000 people.[7] Bardi (spoken by less than 30 people[8]), Yakima Sahaptin (with only 25 speakers[9]), and Chintang (with less than 5000 speakers[10]) are all classified as local languages. The way that a Matrix grammar affects a standardized language could be different from a local one. Take the Bardi people, for example, for whom English has become the more dominant language in the community. Most of the remaining speakers are elders or older adults who are actively teaching the language to younger members who did not natively learn Bardi from birth. Technology that comes out of a Bardi Matrix grammar, possibly in the form of teaching resources, could affect how the language is taught or motivate students to learn. This differs from a language like Meitei, whose speakers learn the language from birth by interacting with members of their community as they grow up. Technology as a result of a Meitei Matrix grammar might be from companies looking to add support for a new language in their app or software, with the goal of providing the Meitei community with the option of using their technology with their own language.

Another way in which a Matrix grammar can affect a language is in relation to technology accessibility. As mentioned previously, there have been issues when it comes to access to technology (i.e. the Internet, computers, specific programs or software) in indigenous communities. Take the Bardi example from above. While the students might be able to use the technology, it may be inaccessible to the elders. This could discourage them from using it or having their students use it and be a waste of time and resources. The same applies to the Meitei example. Although an app or software has Meitei language support, this doesn't mean that people would actually use it, whether that is because the technology is not useful to them or because it is inaccessible.

[7] According to a 2011 census (Eberhard, Simons & Fennig 2023).
[8] According to a 2005 census (Eberhard, Simons & Fennig 2023).
[9] According to a 2007 census (Eberhard, Simons & Fennig 2023).
[10] According to a 2011 census (Eberhard, Simons & Fennig 2023).

6 Value Scenarios

In the following three value scenarios, I use the VSD concerns listed in Section 3.4 to guide the scenarios and their analyses. The concerns covered below include implications for stakeholders, key values, indirect impacts, and systemic effects. The first two value scenarios focus on highlighting the consequences of implementing technology without considering what speakers and communities want and without respecting their wishes. The third value scenario focuses on positive impacts of linguist involvement and respecting the wishes of speakers and communities.

6.1 Yakima Sahaptin Language Education Tool

Scenario Blake is a white American student who just finished taking LING 567. During the class, they worked with Yakima Sahaptin data and a corresponding linguist-created resource grammar to construct a Matrix grammar for the language. They know that there are current efforts within the Yakima community to reclaim the language and culture (Heritage University 2019). Blake has the idea to start a project and create an education tool for students learning Yakima Sahaptin. They want to use a gamified approach and create a computer program similar to Duolingo's language-learning application[11]. This program would allow students to compete with their peers in a fun and interactive format, which Blake hopes motivates them to learn. Currently, there are no permission or data-use restrictions for this language, so Blake believes, based on the information provided by the Grammar Matrix and AGGREGATION project, that they would not need to get permission from the Yakima community (or the linguist who provided the data and/or resource grammar) to create such a program. Blake, knowing they are not an expert in Yakima culture, decides to just omit culture from the examples and exercises in their program and focus on ones that highlight the grammar of the language. They do this by basing the examples and exercises they design on common English sentences that they are familiar with. Blake translates these into Yakima Sahaptin, without considering whether these sentences are confusing or contain concepts that don't make sense within the context of Yakima culture. For example, in some of the sentences Blake references meals they normally eat, which might not line up with meals that community members eat. When the program is finished, Blake shows it to Alex, a teacher at the Yakama Nation Tribal School. Alex decides to try incorporating it into her language classes, which contain high school students (between the ages of fourteen and eighteen). After the program has been used for a couple of months, Blake asks Alex for feedback on the program. Alex describes that, although the students enjoy the gamified and competitive aspects of the program, the example sentences feel unnatural. She has noticed inaccuracies in many of the example sentences, which are rooted in the assumptions that Blake made when designing the program. Alex informs Blake that she is unsatisfied with the program as it is and will not be continuing to use it in her classes.

Analysis There are some glaring issues with Blake's project idea, specifically in relation to VSD concerns of implications for stakeholders, key values, and indirect impacts. Firstly, they assume that members of the Yakima community want such a tool, which relates to the key value of sovereignty (the Yakima people should have sovereignty over their own language and culture). Alex never asked for an education tool but decided to give it a try because she thought it could be

[11] `https://www.duolingo.com/`

beneficial. However, once she realized that the program Blake created goes against another value of the community, one which focuses on language and culture reclamation, she was no longer interested. She is now more skeptical about technology like Blake's program. Somebody could design a tool that is exactly what she wants but, because of how poorly it went with Blake, she might not want to give it a chance, which was not Blake's intention.

Secondly, Blake assumes that language and culture can be separated. However, language and culture are very closely intertwined (as mentioned by Sahaptin language professor Greg Sutterlict in Heritage University (2019)). Real world examples of language will reflect the culture of the speakers of that language. In trying to separate culture from language, Blake makes incorrect assumptions about the culture of speakers of Yakima Sahaptin, which could have adverse affects. The example described in the scenario involved sentences containing references to meals Blake would be accustomed to eating as a white American person. Although quite a bit of American culture has integrated into parts of Yakima life, traditional Yakima foods are not found in American culture. Blake fails to recognize their colonizing viewpoint, assuming that their traditions are correct and that everybody should understand them. Another example is that students using this tool might learn incorrect things about their own community or alternatively be confused or offended by the implications of some sentences. Since the goal of language reclamation is members of the community reclaiming both their language and culture, it is important that tools or programs being created from language data are representative of the culture of the community and the desires of the speakers.

6.2 Matrix/ML Multilingual Translation Tool

Scenario Briar, an NLP developer, wants to use a combination of Matrix-produced grammars and ML technology to create a translation tool. He decides to start by using the Matrix to create grammars for English and Spanish, since he is a native speaker of both and therefore feels more comfortable working with them as opposed to a language he doesn't know. However, his end goal is to have this tool work with other languages as well. Briar finds and uses English and Spanish resource grammars while developing the Matrix grammars and, whenever he is faced with a grammatical implementation decision that is not answered by the resource grammar, he picks the one that is easier to implement. Once he feels the grammar is complete for basic sentences, he begins to integrate ML models to build his translation tool. Briar gets to the point where he believes that his tool is fairly good at translating between simple English and Spanish sentences. Now, he wants to expand it to work with more languages and more complex sentences, so he turns to the grammar engineering community to get access to more Matrix-produced grammars. He expects that members of this community will just hand over their grammars so that Briar can use them in his translation tool. He quickly discovers that they will do no such thing, so he decides to hire a small team of computer scientists and NLP developers (all native English speakers) to help him implement more languages. Within a year, Briar and his team add support for three more languages to the translation tool – German, French, and Italian – using the same methods that he used when implementing English and Spanish. He feels that the tool is ready to be fully released to the public so he focuses on advertising. At first, everything seems to be going swimmingly. Slowly, reviews start coming in from users about how the tool isn't that good. Then, he starts to see backlash from the grammar engineering community that does not agree with his methods. The project is quickly shut down.

Analysis There can be consequences down the line (longer-term use under VSD concerns) when using ease of implementation as the reason for doing something one way over another. Sometimes there really is no way to make a well-informed choice and ease of implementation is a valid reason. However, Briar and his team are relying completely on resource grammars. It is difficult for them, as non-linguists, to have all the tools to make fully-informed choices. If Briar and his team were to consult a linguist when faced with a decision that is not clear-cut based on the resource grammar, they would produce more linguistically-sound grammars. When looking to expand the translation tool, Briar makes a good decision to turn to grammar engineers. However, when he is initially turned down, he should have tried to work with the grammar engineering community and got to the root of why they didn't want to just hand over the grammars they have spent years working on. They would have been able to point Briar towards useful resources that would help introduce him to the best practices in this field. They could also help Briar get in contact with the communities of the languages he wants to add support for in his translation tool, in order to make sure that the members of these communities actually want or need such a tool.

Another VSD concern that is relevant in this scenario is indirect impacts and systemic effects, which becomes applicable when considering the languages that Briar's translation tool supports. English, Spanish, German, French, and Italian are all standardized Indo-European languages. These languages are some of the most commonly-studied languages and scholars have already created better-developed and better-implemented grammars for many of these languages. By choosing to focus on these languages, Briar continues to perpetuate the idea that commonly-spoken languages are the only ones worth creating resources for, something that he probably didn't intend. Briar could have benefited by investing his time and money into building his tool for languages that don't already have a plethora of translation tools available.

6.3 LING 567 Student Grammar

Scenario Bailey and Erin are students in 567 who are trying to build a grammar for Weklati, a made-up language[12] with approximately 150 speakers. When they first pick this language to work with, they immediately contact the linguist who provided the data and whose resource grammar (which is written in English) they will be using – Gael. Bailey and Erin inform Gael of what they are doing and ask if Gael could be their point of contact for clarification questions and grammatical implementation questions that are not answered by the resource grammar. They also ask if there is a way for them to be put in contact with a member of the community that would be interested in what they are doing. Gael replies saying that he is willing to answer questions and that he will reach out to his Weklati contacts that he thinks might be interested. Due to the fast-paced nature of the class, Bailey and Erin must be extremely on top of their work for the class. For each weekly assignment, they start as soon as they can to look over what they need to implement and to prepare questions for Gael. Luckily, Gael is a fairly fast correspondent so they are able to get most of their questions answered before the assignment is due. When this is not possible, they do their best to make implementation choices based on the resource grammar and their prior knowledge and experience (which they slowly add to throughout the duration of the class), using ease of implementation as their last resort. Then, upon hearing back from Gael, if their implementation does not align with his advice, they make changes. Gael eventually gets back to them with the contact information of

[12]I am not using a real language nor a real linguist since I do not want to make assumptions on the behalf of that linguist, or the speakers/community that speak the language.

a speaker – Naomi – who is interested in learning about their project. Luckily, Naomi also speaks English so they are able to email her directly and explain what they are doing. When Naomi replies, she explains how her mother had always been passionate that Weklati be as important as English or German when it comes to being studied. She asks that Bailey and Erin keep her updated on their work. When the class is finished, Bailey decides that she wants to continue working with Weklati language data for her master's thesis (Erin has already decided on a different topic). She reaches out to Naomi and expresses how she wants to continue her work and make a tool that would be useful to Naomi or other members of the Weklati community, providing ideas for possible tools. Naomi replies stating that nobody in the community is looking for a tool at this time. Bailey thanks her for all her correspondence, beginning to brainstorm for new thesis topics.

Analysis Two VSD concerns that are applicable in this scenario are implications for stakeholders and indirect impacts. Bailey, Erin, and Gael are direct stakeholders while Weklati speakers, including Naomi, are indirect stakeholders. Although Bailey and Erin communicated with Naomi, she did not have a direct role in the project and therefore could not be considered a direct stakeholder. Bailey and Erin did exactly what they should in reaching out to Gael, and eventually Naomi, as this set up their respect for where they got their data from – the speakers and communities. Despite this being a 10-week school project that appears, on the surface (or to students just trying to pass a class), to not have real-world implications, they did their best to be respectful of the Weklati community. When Naomi told Bailey that the Weklati community wasn't looking for a language tool, Bailey could have still made one, purely for educational purposes. However, she respected that the Weklati language, and its data, belongs to its community so she decided to find a new thesis topic.

Indirectly, because Bailey and Erin were so committed to having linguist input in their project, they show that it is possible to have linguist involvement in a LING 567 grammar. It is important to acknowledge that Gael was able to respond to Bailey and Erin in a timely manner and that might not always be the case. The linguist might not respond at all or not fast enough to be conducive for such a fast-paced project (although even in the latter case, changes can be made after-the-fact to further improve the grammar). However, it sets the precedent that students can at least try to get feedback from a linguist and, more generally, for better respecting communities.

As mentioned in the scenario, the resource grammar that Bailey and Erin are using is written in English, which means that the example IGT they have use English glosses. This is fairly common when using the Grammar Matrix to build a grammar and is almost always the case for grammars built in LING 567. The view of English as the "default" language is quite prevalent in many fields, but especially within technology- and language- based ones. This view often results in systems that are not as generalizable as they may seem or claim to be (Bender 2019), which connects to the VSD concerns regarding indirect impacts and systemic effects as these types of systems further perpetuate the issue of treating English as the "default" language. In a similar vein, the AGGREGATION project (which has at times been a source of data for LING 567) also uses English glosses however, this is not stated anywhere in the project's documentation. Explicitly stating the glossing language for each language can help to destigmatize the view of English as the "default" language.

7 Discussion

Based on the identification of direct and indirect stakeholders, as well as the description of value scenarios that highlight some of the key concerns that are raised when developing technology, improvements should be made to the AGGREGATION project, the Grammar Matrix, and to LING 567 so that the stakeholders are better considered. The initial step is to record glossing languages in documentation, even (especially) when it's English. For the AGGREGATION project, the glossing language should be included for each dataset. Grammars produced by the Grammar Matrix do not have documentation in a similar way to the AGGREGATION project, but the choices file for a grammar should, in addition to the language being studied, include the glossing language (or languages) being used. Although the current curriculum of LING 567 doesn't require students to write papers about their projects, students often use some directory or repository with version control to track their implementation of a grammar, which usually includes some front-facing documentation file to introduce the project. This would be a good place for the students to include the glossing language. Additionally, for students that extend their work in LING 567 to other projects that do involve a written paper, the glossing language should be clearly stated within said paper.

The next step would be to contact the communities and linguists for the languages in the repository, especially those that have no permission restrictions listed, in order to start building a relationship with the communities and to ascertain what they want to happen with their data. Within the context of LING 567, students should, when possible, contact the linguist who provided the data for the language they are working on. This should happen in the early stages of the project, ideally one of the very first things. The initial contact with the linguist should give context for the class and project, how the data is being used, the possible implications, and options for data permission restrictions. The professor might find it useful to create an email template with some of this information that the students can modify to fit their specific project. In doing this, the students would confirm that the linguist is aware that the data is being used for the class and that they agree with how the data is being used (or if they don't agree, alternative options are provided). After the initial contact, follow-up communication could connect the students with speakers of the language they are working with to determine their wishes with respect to the data and how it is being used. A similar process and email template could be used for the AGGREGATION project, where the linguist for each dataset is contacted, as well as the community for each language. This could be done by the person in charge of the project, be delegated to a student as paid work, or could be part of a student's research or project (e.g. a thesis that focuses on documentation).

Additionally, it would be in the best interest of the students of LING 567 to consider the possible impacts of their work – who it will impact and how it might affect them – especially outside of the classroom setting. They should try and incorporate contact with communities and linguists into their projects because it would instill good language-data-use habits and produce a more linguistically-sound grammar. For grammar engineers that have been Matrix developers for some time, this paper will hopefully encourage them to reach out to speakers, communities, and/or linguists in order to start a line of communication about their work.

As for projects that exist outside of the academic scope (i.e. outside of projects like the AGGREGATION project and the Grammar Matrix or classes like LING 567), there isn't much that can be done to explicitly improve how they handle data and how they treat the people they got it from. However, raised awareness and better documentation (that is clear and easy to find) can increase

people's respect towards languages, communities, and their speakers. This can set a precedent, encouraging further improvements.

8 Ethical Considerations

Since much of what is discussed relates to student work, it is important to recognize that it is for a class. Most of what the students work on will not be used in any contexts outside of the classroom. The goal is to learn how to use the tools and to get practice with them. However, it is still important that good habits and practices start at the beginning, especially since there are cases of students building off of their work from that class in further academic contexts that could affect stakeholders or have other societal impacts.

9 Conclusion

Using value sensitive design methods, I was able to explore the societal impacts and ethical implications of the AGGREGATION data repository, as well as the Grammar Matrix and how it is used in LING 567. Through stakeholder analysis, I was able to identify students, grammar engineers, and linguists as direct stakeholders and speakers, communities, and linguists as indirect stakeholders. With the value scenarios I described, I showed how these stakeholders should be better incorporated into the grammar engineering process and how the focus needs to be on the speakers, communities, and what they want, especially when it comes to creating technology for a community. Additionally, I showed how the AGGREGATION data repository needs to be better documented.

In the future, it would be interesting to see the improvements I suggested incorporated into LING 567 and see how it affects the class. Would linguist or speaker involvement be too slow to effectively work within the context of a 10-week-long class? Would such involvement lead to better grammars? Also, further work in relation to VSD could be made with different methods than discussed here. For example, looking at the co-evolution of technology and social structures, performing value-oriented semi-structured interviews, or conducting an ethnographically informed inquiry regarding values and technology. These could provide new insights in other parts of the grammar engineering process than focused on in this paper.

Acknowledgements

I would like to thank Emily Bender in particular for her feedback during the process of writing this paper. Additionally, I would like to thank Clara Brandt, Elizabeth Conrad, Gita Dhungana, and Elizabeth Snell Okada for their feedback.

References

Bender, Emily M. 2019. The #BenderRule: On naming the languages we study and why it matters. *The Gradient*.

Bender, Emily M., Joshua Crowgey, Michael Wayne Goodman & Fei Xia. 2014. Learning grammar specifications from IGT: A case study of Chintang. In *Proceedings of the 2014 workshop on the use of computational methods in the study of endangered languages*, 43–53.

Bender, Emily M., Scott Drellishak, Antske Fokkens, Laurie Poulson & Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation* 8(1). 10.1007/s11168-010-9070-1, 23–72. https://doi.org/10.1007/s11168-010-9070-1.

Bender, Emily M. & Guy Emerson. 2021. Computational linguistics and grammar engineering. *Head-Driven Phrase Structure Grammar: The handbook*. 1105–1153.

Bender, Emily M., Dan Flickinger & Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk & Richard Sutcliffe (eds.), *Proceedings of the workshop on grammar engineering and evaluation at the 19th international conference on computational linguistics*, 8–14. Taipei, Taiwan.

Bender, Emily M., Dan Flickinger, Stephan Oepen, Woodley Packard & Ann Copestake. 2015. Layers of interpretation: On grammar and compositionality. In Matthew Purver, Mehrnoosh Sadrzadeh & Matthew Stone (eds.), *Proceedings of the 11th international conference on computational semantics*, 239–249. London, UK: Association for Computational Linguistics. https://aclanthology.org/W15-0128.

Bender, Emily M., Michael Wayne Goodman, Joshua Crowgey & Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the 7th workshop on language technology for cultural heritage, social sciences, and humanities*, 74–83.

Bird, Steven. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, 3504–3519.

Bird, Steven. 2022a. Beyond technological solutions: How we create a world that sustains its languages. *Language Technologies and Language Diversity Tecnologies de la llengua i diversitat lingüística*. 167–173.

Bird, Steven. 2022b. Local languages, third spaces, and other high-resource scenarios. In Smaranda Muresan, Preslav Nakov & Aline Villavicencio (eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 7817–7829. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.539.

Bond, Francis, Stephan Oepen, Eric Nichols, Dan Flickinger, Erik Velldal & Petter Haugereid. 2011. Deep open-source machine translation. *Machine Translation* 25. 87–105.

Copestake, Ann. 2002. *Implementing typed feature structure grammars*. Vol. 110. CSLI publications Stanford.

Copestake, Ann. 2007. *Natural Language Processing*. University Lecture. https://www.cl.cam.ac.uk/teaching/0809/NLP/lectures.pdf.

Crysmann, Berthold & Woodley Packard. 2012. Towards efficient HPSG generation for German, a non-configurational language. In *Proceedings of COLING 2012*, 695–710.

Dobrin, Lise, Peter K. Austin & David Nathan. 2014. Dying to be counted: The commodification of endangered languages in documentary linguistics. *Language documentation and description* 6.

Dyson, Laurel Evelyn, Max Hendriks & Stephen Grant. 2006. *Information technology and indigenous people*. IGI Global.

Eberhard, David M., Gary F. Simons & Charles D. Fennig (eds.). 2023. *Ethnologue: Languages of the world*. Twenty-sixth. Dallas, TX, USA: SIL International. http://www.ethnologue.com.

Fostar, Jonathan Blake. 2021. Like death but without death: The language-death-metaphor and another option. *Linguaculture* 12(2). 85–101.

Friedman, Batya, David G. Hendry & Alan Borning. 2017. A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction* 11(2). 63–125.

Georgi, Ryan. 2016. *From Aari to Zulu: Massively multilingual creation of language tools using interlinear glossed text* dissertation.

Goodman, Michael Wayne. 2018. *Semantic operations for transfer-based machine translation*. University of Washington dissertation.

Goodman, Michael Wayne, Joshua Crowgey, Fei Xia & Emily M. Bender. 2015. Xigt: extensible interlinear glossed text for natural language processing. *Language Resources and Evaluation* 49. 455–485.

Heritage University. 2019. *Saving the language*. https://www.heritage.edu/saving-the-language/.

Howell, Kristen & Emily M. Bender. 2022. Building analyses from syntactic inference in local languages: An HPSG grammar inference system. In *Northern european journal of language technology, volume 8*.

Howell, Kristen, Emily M. Bender, Michel Lockwood, Fei Xia & Olga Zamaraeva. 2017. Inferring case systems from IGT: Enriching the enrichment. In *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages*, 67–75.

Krieger, Hans-Ulrich & Ulrich Schafer. 1994. TDL-A type description language for constraint-based grammars. In *COLING 1994 volume 2: the 15th International Conference on Computational Linguistics*. https://aclanthology.org/C94-2144.

Leonard, Wesley Y. 2008. When is an "extinct language" not extinct. *Sustaining linguistic diversity: Endangered and minority languages and language varieties*. 23–33.

Nathan, Lisa P., Predrag V. Klasnja & Batya Friedman. 2007. Value scenarios: A technique for envisioning systemic effects of new technologies. In *Chi '07 extended abstracts on human factors in computing systems* (CHI EA '07), 2585–2590. San Jose, CA, USA: Association for Computing Machinery. https://doi.org/10.1145/1240866.1241046.

Oepen, Stephan. 2001. {*incr tsdb ()*}*–Competence and performance laboratory*. Tech. rep. User Manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken.

Packard, Woodley. 2015. *Full forest treebanking* dissertation.

Pollard, Carl & Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Ramponi, Alan. 2024. Language Varieties of Italy: Technology Challenges and Opportunities. *Transactions of the Association for Computational Linguistics* 12. 19–38. https://doi.org/10.1162/tacl_a_00631.

Smith, Linda Tuhiwai. 2021. *Decolonizing methodologies: Research and indigenous peoples*. Bloomsbury Publishing.

Taiuru, Karaitiana. 2023. Māori data is a taonga. *Indigenous Research Design: Transnational Perspectives in Practice*. 173–192.

Toth, Katalina, Daisy Smith & Daphne Giroux. 2018. Indigenous peoples and empowerment via technology. *First Peoples Child & Family Review* 13(1). 21–33. https://doi.org/https://doi.org/10.7202/1082388ar.

Vauquois, Bernard. 1968. Structures profondes et traduction automatique: Le systeme du ceta.

Wax, David Allen. 2014. *Automated grammar engineering for verbal morphology* dissertation.

Zamaraeva, Olga, Chris Curtis, Guy Emerson, Antske Fokkens, Michael Wayne Goodman, Kristen Howell, T.J. Trimble & Emily M. Bender. 2022. 20 years of the grammar matrix: cross-linguistic hypothesis testing of increasingly complex interactions. *Journal of Language Modelling* 10(1). 49–137.

Zamaraeva, Olga, Kristen Howell & Emily M. Bender. 2019. Handling cross-cutting properties in automatic inference of lexical classes: A case study of Chintang. In *Proceedings of the 3rd workshop on the use of computational methods in the study of endangered languages volume 1 (papers)*, 28–38.

Zamaraeva, Olga, František Kratochvíl, Emily M. Bender, Fei Xia & Kristen Howell. 2017. Computational support for finding word classes: A case study of Abui. In *Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages*, 130–140.