Inferring Case Systems from IGT:
Impacts and Detection of Variable Glossing Practices

Kristen Howell[*], Emily M. Bender[*], Michael Lockwood[*], Fei Xia[*], and Olga Zamaraeva[*]

[*]Department of Linguistics, University of Washington, Seattle, WA, U.S.A.
{kphowell, ebender, lockwm, fxia, olzama}@uw.edu

Abstract

In this paper, we apply two methodologies of data enrichment to predict the case systems of languages from a diverse and complex data set. The methodologies are based on those of Bender et al. (2013), but we extend them to work with a new data format and apply them to a new dataset. In doing so, we explore the effects of noise and inconsistency on the proposed algorithms. Our analysis reveals assumptions in the previous work that do not hold up in less controlled data sets.

1 Introduction

This work is situated within the AGGREGATION Project whose aim is to facilitate analysis of data collected by field linguists by automatically creating computational grammars on the basis of interlinear glossed text (IGT) and the LinGO Grammar Matrix customization system (Bender et al., 2010). Previous work by the AGGREGATION Project has looked at answering specific, high-level typological questions for many different languages (Bender et al., 2013) as well as answering as much of the Grammar Matrix customization system's questionnaire as possible for one specific language (Bender et al., 2014). In this paper, we revisit the case-related experiments done by Bender et al. (2013) in light of new systems for standardizing and enriching IGT and using a broader data set. Specifically, where Bender et al. considered only data from small collections of IGT created by students in a grammar engineering class (Bender, 2014), we will work with the larger and more diverse data sets available from ODIN version 2.1 (Xia et al., 2016). These data sets contain a great deal of noise in terms of inconsistent glossing conventions, missing glosses and data set bias. However, while these data sets are noisier, they do benefit from more sophisticated methods for enriching IGT (specifically projecting structure from the English translation line to the source language line; Georgi (2016)). Additionally, we explore cleaning up some noise in the glossing, using the Map Gloss methodology of Lockwood (2016).

In the following sections, we provide a description of the methodology and data sets we build on (§2), before laying out the methodology as developed for this paper (§3) and presenting the numerical results (§4). The primary contribution of our paper is in (§5), where we do an error analysis and relate our results to sources of bias.

2 Background

This section briefly overviews the previous work that we build on in this paper: methodology developed by the RiPLes and AGGREGATION projects to extract typological information from IGT (§2.1), the construction and enrichment of the ODIN data set (§2.2), and Map Gloss system for regularizing glosses in IGT (§2.3).

2.1 Inferring Case Systems from IGT

Bender et al. (2013) began work on automatically creating precision grammars on the basis of IGT by using the annotations in IGT to extract large-scale typological properties, specifically word order and case system. These typological properties were defined as expected by the Grammar Matrix customization system (Bender et al., 2010), with the goal of eventually providing enough information (both typological and lexical) that the system can automatically create useful implemented grammars. A proof of concept for this end-to-end system was conducted with data from Chintang in Bender et al. (2014).

In their case experiment, Bender et al. (2013) explored two methods for inferring case systems from IGT. The first, called GRAM, counted all of the case grams for a particular language and ap-

Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 67–75,
Honolulu, Hawai'i, USA, March 6–7, 2017. ©2017 Association for Computational Linguistics

plied a heuristic to determine case system based on whether or not certain grams were present (NOM, ACC, ERG, ABS). The second method, called SAO, collected all of the grams on all intransitive subject (S), transitive subject (A) and transitive object (O) NPs, and then proceeded with the assumption that the most common gram in each role was a case gram. The grammatical role was determined by mapping parses of the English translation (using the Charniak parser (Charniak, 1997)) through the gloss line onto the source language line, according to the methodology set forth in the RiPLes project for resource-poor languages (Xia and Lewis, 2007). Because this method looks for the most frequent gram to identify the case marking gram, it is not essential that the gloss line follow a specific glossing convention, provided that the grams are consistent with each other. As a result, this method was expected to be suited to a wider range of data than GRAM.

The data for this experiment comprised 31 languages from 17 different language families. Data was developed in a class in which students use descriptive resources to create testsuites and Grammar Matrix choices files (files that specify characteristics of the language so the Grammar Matrix can output a starter grammar) and has now been curated by the LANGUAGE CoLLAGE project (Bender, 2014). Because the testsuites were constructed for grammar engineering projects, data from the testsuites are not representative of typical language use nor of typical field data collections, but nonetheless illustrate grammatical patterns. In addition, the testsuites generated in this class conform to standard glossing procedures (specifically the Leipzig Glossing Rules (LGR; Bickel et al. (2008)) and are annotated for the phenomena they represent.

The results of Bender et al. (2013) are given in Table 1, where the baseline is a ‘most frequent type’ baseline, i.e. chosen according to the most frequent case system in a typological survey (here, this would be neutral aka no case).[1]

In this experiment, both GRAM and SAO performed better than the baseline. The GRAM method outperformed SAO, which was attributed to the small size and LGR-compliant nature of the testsuites.

Data Set	Number of languages	GRAM	SAO	Baseline
DEV1	10	0.900	0.700	0.400
DEV2	10	0.900	0.500	0.500
TEST	11	0.545	0.545	0.455

Table 1: Accuracy of case-marking inference as reported in (Bender et al., 2013)

2.2 ODIN Data Set

The Online Database of Interlinear Text (Lewis and Xia, 2010; Xia et al., 2016) was developed by crawling linguistics papers and extracting IGT. In ODIN 2.1, this data is enriched with dependency parses by parsing the translation line with the MSTParser (McDonald et al., 2005), aligning the translation and language lines, and then projecting the syntactic structure of the English parse onto the language line, according to the methodology set forth by Georgi (2016). Dependency parses make explicit the grammatical role of items in a sentence. For our purposes, this means that identifying subjects, agents and objects is more straightforward than it was for Bender et al. (2013), who were working with projected constituency structures.

2.3 Map Gloss

One issue that arises when working with IGT, especially IGT taken from multiple different sources (as is found in the ODIN collection), is that the glossing cannot be assumed to be consistent. Different authors follow different glossing conventions, including different sets of grams; grams may be misspelled; and glossing may be carried out at different levels of granularity. Map Gloss (Lockwood, 2016) was developed to address the first two of these sources of variation, which are also found in the LANGUAGE CoLLAGE data (though to a lesser extent). Map Gloss takes in a set of IGT for a language and outputs a standard set of grams for that language, as well as a mapping from glosses observed in the IGT to the standard set. Lockwood (2016) constructed a gold standard set of grams that follows the Leipzig Glossing Rules (Bickel et al., 2008) and the GOLD Ontology (Indiana University, 2010) conventions.

Map Gloss maps common misspelled glosses to their correct form, normalizes glosses such as IMP (for ‘imperfective’) to less ambiguous forms such as IPFV, adds grams where they were left out such as finding *he* in the gloss line instead of 3SG.M,

[1] We take this heuristic from Bender et al. (2013), although when we applied this heuristic to the 2013 data, our results did not quite match those in Table 1.

and splits grams that were combined such as 3SGM to 3SG.M. Finally, Map Gloss allows for grams that are language specific (i.e. not known to the gold standard set but also not targets for correction). Though the gloss line in IGT will typically mix lemmas and grams, and may sometimes also contain part of speech tags, the final normalized set that Map Gloss outputs for a language will contain only the grams.

3 Methodology

In this experiment, we extend and adapt the methodology of Bender et al. (2013), designed to work with projected constituency structures, to use the dependency structures available for the enriched IGT in ODIN 2.1. To accomplish this, we first reimplemented the software from Bender et al. (2013) to work with the Xigt format (Goodman et al., 2015). Xigt, beyond being the format used in ODIN 2.1, has many advantages for this kind of work because it is set up to directly encode not only the base IGT but also further annotations (enrichments) over that IGT (in a stand-off fashion), such that one can easily query for items such as the subject in the projected dependency structure.

For better comparability with the results reported by Bender et al. (2013), ideally we would be working with the same languages. However, some of the languages used in the 2013 experiment have few or no instances of IGT in ODIN. These languages were not expected to yield useful results, so we added eight new languages which were available in both LANGUAGE COLLAGE and ODIN. Our final data set comprises 39 languages from 23 distinct language families, as shown in Table 3 below. The quantity of data varies widely across these languages, as do the number of authors contributing the data. The IGT for French [fra], for example comes from 4,787 separate documents, some of whom have overlapping authors. To account for the added languages, we re-ran the scripts from Bender et al. (2013) on their original data as well as the LANGUAGE COLLAGE testsuites for the newly added languages (shown in the LANGUAGE COLLAGE lin of Table 4).

We used Map Gloss (Lockwood, 2016) to standardize the glosses from the ODIN data. ODIN includes IGT from a wide range of authors with their own conventions for gloss naming. We hypothesized that mapping these glosses to a standard set of grams is important in helping our inference procedure correctly assign case systems. Map Gloss features a default set of standard glosses and a generic training set, which we used, in lieu of annotating our own training set. For this experiment we were chiefly interested in the grams for case, which Map Gloss handles quite well off the shelf, using a list of commonly used case glosses that map to a standard set for each language. For example INS, INST, INSTR and INSTRUMENTAL all map to INS. Map Gloss was run on each of the ODIN language files individually to identify a standard set of grams. We then generated new Xigt files for each language, replacing the existing case glosses with the standardized case glosses from Map Gloss.

For case system extraction we adapt both the GRAM and SAO methodologies set forth in the 2013 experiment. The GRAM method loops through the IGT collecting glosses from a set of licensed grams (or tags) associated with case. We assign case systems according to the presence or absence of NOM, ACC, ERG and ABS, as in Table 2. This methodology assumes compliance with the Leipzig Glossing Rules, and therefore its performance relies on data being glossed according to these conventions with regards to case.

Case	Case grams present	
system	NOM ∨ ACC	ERG ∨ ABS
neutral		
nom-acc	✓	
erg-abs		✓
split-erg	✓	✓

Table 2: GRAM case system assignment rules (Adapted from Bender et al. (2013))

The second method, SAO, is intended to be less dependent on glossing choices. This method uses the dependency parses in ODIN to identify the subject of intransitive verbs (S), the agent of transitive verbs (A) and the patient of transitive verbs (O). We consider only clauses that appear to be simple transitive or intransitive clauses, based on the presence of an overt subject and/or object, and collect all grams for each argument type. We assume the most frequent gram for each argument type (S, A or O) to be the case marker. We then use the following rules to assign case system, where S_g, O_g and A_g denote the most frequent grams associated with these argument positions:

- Nominative-accusative: S_g=A_g, and $S_g \neq O_g$
- Ergative-absolutive: S_g=O_g, and $S_g \neq A_g$
- Neutral: S_g=A_g=O_g, or $S_g \neq A_g \neq O_g$ and S_g, A_g, O_g also present on each of the other argument types
- Tripartite: $S_g \neq A_g \neq O_g$, and S_g, A_g, O_g (virtually) absent from the other argument types
- Split-ergative: $S_g \neq A_g \neq O_g$, and A_g and O_g are both present in the list for the S argument type

(Bender et al., 2013)

The space of outputs of the two systems differ slightly. The GRAM method predicts four possible case systems: neutral, nominative-accusative, ergative-absolutive and split-ergative. The SAO method is a little more robust: in addition to the four case systems predicted by GRAM, it can also predict the tripartite case system. We compare the predicted case systems from each system to a gold standard, collected from the choices files and notes included with the LANGUAGE CoLLAGE data. In the 2013 experiment, the gold standard for case systems was taken from the choices files produced by the grammar engineering students who produced the original testsuites. The Grammar Matrix customization system (Bender et al., 2010) allows users to choose from a list of possible case systems (as well as other linguistic phenomena) and records this information in a file named choices. However, in some cases, the student might not have used the customization system to establish their case system, so in the present experiment we reviewed their notes for clarification if no case system was specified in the choices file. Because our methodologies predict split-ergativity but are not so refined as to specify subtypes of split-ergativity which are available in the Grammar Matrix customization system, we have included both subtypes 'split-v' (a split based on properties of the verb) and 'split-s' (a split based on properties of the subject) as part of the super-type 'split-ergative' for evaluation.

4 Results

We re-ran the 2013 experiment on all of the data from the original experiment (DEV1, DEV2, and TEST) in addition to LANGUAGE CoLLAGE testsuites for the eight new languages. Both GRAM and SAO were run on ODIN data sets for each language both before and after running Map Gloss to standardize the case grams. In addition, we generated the baseline value in the same manner as Bender et al. (2013), using the 'most common type' heuristic. According to Comrie (2011) this is again 'neutral'. The results are given in Table 4.[2]

Our results for LANGUAGE CoLLAGE are notably lower than the results reported in Bender et al. (2013), as shown in Table 1. This can be attributed to two changes from the 2013 experiment. First, we added 8 new languages which were not included in the 2013 results. Second, we updated the gold standard for some of the original languages, after mining through the grammar engineers' notes, rather than merely relying on the case system identified in the choices files. As a result, some languages now have specified case systems in the gold standard, which were assumed to be 'none' (or neutral) in the original experiment.

Although Map Gloss did standardize the case grams in our experiment, it had no measurable effect on the results of the experiment. Map Gloss changed 2.2% of case grams across data sets. The percentage of case grams changed per data set ranged from 0% for many to 50%. For example, in the Hausa data, 30% of case grams were changed, standardizing *sub* and *sbj* to *subject*. However, the most commonly found subject gram in the Hausa data was *abdu* and the most common object gram was *of*. This sort of data set bias is discussed in more detail in section 5.4 but it demonstrates the small impact that Map Gloss made on the results even when it made numerous changes to the data. For many other languages, no case grams were changed by Map Gloss because none were inconsistently spelled or misspelled, or those case grams that were inconsistently or misspelled were not in the subset relevant to either GRAM's or SAO's heuristics (NOM, ACC, ERG, ABS for GRAM or the most frequent gram for SAO).

The accuracy of SAO on ODIN data was just below baseline and GRAM preformed only slightly better with accuracy rates of 41.0% and 56.4% respectively. These results demonstrate that the ODIN data presents an even greater challenge than

[2] Code, data and instructions for reproducing the results in Table 4 can be found at `http://depts.washington.edu/uwcl/aggregation/ComputEL-2-Case.html`

[3] The baseline result is the same across the data sets, because the baseline compares the gold standard to the 'most common' case system, and is independent of the dataset.

Language Name	ISO	Language Family	# IGTs	# IGT with Dependency	Gold Standard Case System
French	fra	Indo-European	7412	1322	neutral
Japanese	jpn	Japanese	6665	2484	nom-acc
Korean∗	kor	Korean	5383	2208	nom-acc
Icelandic	isl	Indo-European	4259	1100	neutral
Russian	rus	Indo-European	4164	1579	nom-acc
Hausa	hau	Afro-Asiatic	2504	1085	neutral
Indonesian∗	ind	Austronesian	1699	1075	neutral
Georgian	kat	Kartvelian	1189	463	split-erg
Tagalog	tgl	Austronesian	1039	418	erg-abs
Thai	tha	Thai-Kadai	692	184	neutral
Czech	ces	Indo-European	664	257	nom-acc
Zulu	zul	Niger-Congo	604	86	neutral
Kannada∗	kan	Dravidian	523	300	nom-acc
Chichewa∗	nya	Niger-Congo	477	151	neutral
Old English	ang	Indo-European	431	136	nom-acc
Welsh	cym	Indo-European	404	191	neutral
Vietnamese	vie	Austro-Asiatic	352	176	neutral
Taiwanese	nan	Sino-Tibetan	275	148	neutral
Pashto	pbt	Indo-European	274	98	erg-abs
Tamil	tam	Dravidian	244	90	nom-acc
Malayalam	mal	Dravidian	172	91	nom-acc
Breton	bre	Indo-European	74	50	nom-acc
Lillooet∗	lil	Salishan	72	16	neutral
Ojibwa	ojg	Algic	64	24	neutral
Hixkaryana	hix	Cariban	62	27	neutral
Lushootseed	lut	Salishan	52	16	neutral
Shona	sna	Niger-Congo	50	18	neutral
Huallaga	qub	Quechuan	46	27	nom-acc
Arabic (Chadian)∗	shu	Afro-Asiatic	41	12	nom-acc
Ainu	ain	Ainu	40	21	nom-acc
Ingush	inh	Nakh-Daghestanian	23	13	erg-abs
Arabic (Moroccan)	ary	Afro-Asiatic	14	2	neutral
Haida∗	hdn	Haida	7	1	split-erg
Mandinka	mnk	Mande	3	0	neutral
Hup	jup	Nadahup	2	1	nom-acc
Yughur∗	uig	Altaic	2	0	nom-acc
Jamamadi	jaa	Arauan	1	1	neutral
Sri Lankan Creole Malay	sci	Malay	0	0	split-erg
Bosnian-Serbo-Croatian	hbs	Indo-European	0	0	nom-acc

Table 3: Languages used in our experiment and their IGT counts in ODIN 2.1. Language families are taken from Haspelmath et al. (2008). The asterisk indicates the 8 new languages that were added to the set of 31 languages from the 2013 experiment

Data	GRAM	SAO	BASELINE
LANGUAGE COLLAGE	0.743	0.589	0.462
ODIN	0.564	0.410	0.462
ODIN + MAP GLOSS	0.564	0.410	0.462

Table 4: Prediction accuracy for 39 languages.[3]

the LANGUAGE COLLAGE data for both methods. While these results are modest, they provide valuable insight into both the methodology and the data, which will be useful in future work to those developing inference systems and those who wish to benefit from them.

5 Error Analysis

In this section we report on the results of our error analysis, specifically looking into the likely causes of particular languages being misclassified by each system.

5.1 Little or no data in ODIN

Two of the languages, Sri Lankan Creole Malay [sci] and Bosnian-Serbo-Croatian [hbs], were not present in ODIN and therefore, the system had no data with which to predict the case and defaulted to 'neutral'. Furthermore, ten other languages had fewer than fifty IGTs in the ODIN collection. If we were to remove these twelve languages from the data, the adjusted results would improve marginally, as shown in Table 5.[4]

Data	GRAM	SAO	BASELINE
LANGUAGE COLLAGE	0.889	0.704	0.556
ODIN	0.593	0.481	0.556
ODIN + MAP GLOSS	0.593	0.481	0.556

Table 5: Prediction accuracy for the 27 languages with at least 50 IGTs

5.2 Availability of dependency parses

One of the anticipated benefits of using ODIN 2.1 data was the availability of dependency parses that could be used for the SAO method. These parses identified a subject and direct object, such that the corresponding noun could be extracted and broken into glosses. However, while the presence of these dependency parses is helpful for this type data, only a fraction of the IGTs had a subject and/or object that the dependency structure had successfully identified. However the reduced number of available IGT due to the lack of dependency parses had little affect on SAO. Filtering out data sets with fewer than 50 IGT with dependency parses, SAO's accuracy is 41.0%, which is no improvement over the results in Table 4.

5.3 Absence of Case Grams

Eight of the languages in ODIN (not counting the two for which there was no data) contained no case glosses at all. Of those eight, five had a neutral case system. The other three were Breton [bre] (nom-acc), Chadian Arabic [shu][5] (nom-acc) and Haida [hdn] (split-erg).[6] Twelve other languages had an average of $<$ 1 case gram per IGT. Of these 20 languages, the only case systems which GRAM correctly predicted were those with neutral case systems. SAO preformed comparably on these languages, only correctly predicting those with neutral case systems and Breton [bre] (nom-acc), which was the result of IGT bias, discussed in more detail in section 5.4.

The under-glossing of case grams is symptomatic of linguistics papers that gloss only the distinctions relevant to the argument at hand, rather than giving full IGT. We hope that the Guidelines for Supplementary Materials Appearing in LSA Publications will have an impact on the robustness of IGT glossing in future data collected by ODIN.[7]

5.4 IGT Bias

The vast majority of the predictions made by the SAO method were not based on case glosses at all. Due to a poverty of case glosses in the data, the most common subject, agent and object glosses in the data were usually root nouns. Our algorithm hinges on the hypothesis that case would be the most common gloss across the data if it were glossed on all nouns. We expected that in some languages, person, number or definiteness grams would out-number case grams; however, this was rarely the case in the ODIN data for the 39 languages we sampled. In French [fra] and Welsh [cym], the most common subject and agent gram was *I*, while in other languages it was *he* or 1SG. Breton's most common subject and agent was *children* and most common object was *books*. While in the case of Breton and others, this led to the correct prediction of nom-acc (because the subject and agent were the same as each other and different from the object) the prediction was made for the wrong reasons. Other most-frequent glosses were *dragon*, *jaguar*, *Maria* and *cows*, all of which were so frequently used in their respective data sets that they outnumbered inflectional morphemes that might be more informative for our purposes.

On the one hand, these results demonstrate that our assumption that case grams would be the most common among noun phrases is far too strong when applied to real world data. While the carefully constructed testsuites from LANGUAGE COLLAGE glossed grams throughly and took care to vary their vocabulary, data in linguistics papers may not be so diverse. Indeed in future work, our algorithm should exclude root glosses. Nevertheless, we consider the trend towards using the same noun as the subject across a dataset to be a form of 'IGT bias', as identified by Lewis and Xia (2008). Specifically, it is likely the result of the strategies used for elicitation or the way in which authors chose sentences to include in their

[4] We note that results also improve on the LANGUAGE COLLAGE data set when we restrict our attention to these languages.

[5] Chadian Arabic only expresses case-marking on pronouns.

[6] Haida only expresses overt case on pronouns and is generally considered to have a neutral case system.

[7] http://www.linguisticsociety.org/sites/default/files/Supplemental%20Materials%20Guide%20for%20LSA%20Pubs.pdf, accessed Feb 13, 2017.

papers. While keeping to a restricted range of vocabulary can perhaps be useful for systematic documentation or exposition of particular grammatical phenomena, it can also result in highly biased data sets. For our purposes, more varied sentences would produce more helpful data sets—and we believe that this is true for other research purposes as well. In addition to bias in the words themselves used for annotation, additional bias may be introduced if a linguist is collecting data for a specific phenomenon. For example, if a language includes an unrepresentatively large set of intransitive or unergative verbs, the system might not have data with which to identify a nominative-accusative system. This type of data set bias can be overcome by collecting a diverse set of data from a variety of sources that is large enough to overcome the biases of a particular set (Lewis and Xia, 2008).

5.5 Gold Standard

A final contributor to the accuracy measurements was the state of the gold standard itself. We do not consider this a source of error, but rather an inevitability of working with low-resource languages and the very reason a system such as this is useful. Some of the languages classified as having a neutral case system (corresponding to 'no-case' in the choices files we use as our gold standard) might be better analyzed as in fact having (non-neutral) case systems. The classification in the gold standard, rather than being an assertion on the part of the grammar engineer who created the choices file, might instead indicate that they did not have sufficient evidence to specify a case system. As noted in §4, we did adjust the gold standard away from 'no-case' for some languages, on the basis of the grammar engineers' notes. The cases described here, in contrast, did not have such evidence in the grammar engineers' notes. In some cases, the analyses made by the grammar engineering students that we used to develop our gold standard are not consistent with more common analyses. Icelandic for example was classified 'neutral' in our gold standard, but is widely considered nominative-accusative, as analyzed by Wunderlich Wunderlich (2003) in the ODIN data.

6 Discussion

We acknowledge that the primary result of this work is to show that this is a hard problem to approach automatically. Furthermore, given the fact that in the evaluation there is one data point per language, it is difficult for methods like Map Gloss, which work at the level of improving consistency of glossing of particular examples, to move the needle much. Nonetheless, we think that it is still interesting to pursue methods such as those described here. Aside from the big-picture goal of automatically creating precision grammars and using them to further the analysis of data collected by descriptive and documentary linguists, there is the fact that automated methods can provide interesting summaries of what is found in data sets.

For example, the results of the GRAM method on data collected from a variety of linguists brings to light varying analyses of the language's case system that can prompt a linguist to investigate further. GRAM predicted Lillooet [lil] and Indonesian [ind] to be split-erg. The Lillooet data contained nominative and ergative glosses, suggesting either a split-ergative analysis or authors of IGT-bearing documents choosing different analyses. In fact we find both nominative and ergative glosses in the data from Geurts (2010) and Wharram (2003), suggesting that they have either adopted a split-ergative analysis or that the case system demonstrated some complexity and was not the focus of their work. Indonesian had nominative, accusative and ergative glosses (some in the same IGT), suggesting a split-erg or tripartite analysis for this language as well. In fact the data came from a discussion of ergative-accusative mixed systems by Wunderlich (2006). Thus even in the capacity of mining the case grams used, the GRAM method is useful in shedding light on potential alternative analyses for a given language.

7 Conclusion

We have replicated and extended an experiment designed to automatically predict case system for languages using IGT as part of a larger goal to make inferences about a variety of linguistic characteristics. The results are mixed and in our analysis we identified a number of challenges in working with broad collections of data for low-resource languages. While IGT is a rich source of linguistic information, we find that the information that included in annotation may be incomplete or highly biased. While this is an inevitability in field data which is still in the process of being curated, we as linguists can strive to publish data whose annotation is as complete as possible, given the state of our analysis of the language at the time of publication. Referring again to the Guidelines for Supplementary Materials Appearing in LSA Publications,[8] we strongly encourage our fellow linguists to publish carefully annotated data.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1561833.

References

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, pages 1–50. 10.1007/s11168-010-9070-1.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and 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*, pages 74–83, Sofia, Bulgaria, August. Association for Computational Linguistics.

Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, and 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*, pages 43–53, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Emily M. Bender. 2014. Language CoLLAGE: Grammatical description with the LinGO grammar matrix. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2447–2451, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1508.

Balthasar Bickel, Bernard Comrie, and Martin Haspelmath. 2008. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Max Planck Institute for Evolutionary Anthropology and Department of Linguistics, University of Leipzig.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of AAAI-1997*.

Bernard Comrie. 2011. Alignment of case marking of full noun phrases. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich.

Ryan Georgi. 2016. *From Aari to Zulu: Massively Multilingual Creation of Language Tools using Interlinear Glossed Text*. Ph.D. thesis, University of Washington.

Bart Geurts. 2010. Specific indefinites, presupposition and scope. *Presuppositions and Discourse: Essays Offered to Hans Kamp*, 21:125.

Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender. 2015. Xigt: Extensible interlinear glossed text. *Language Resources and Evaluation*, 2:455–485.

Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors. 2008. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. http://wals.info.

Department of Linguistics (The LINGUIST List) Indiana University. 2010. General ontology for linguistic description (gold). http://linguistics-ontology.org/gold/2010.

William D. Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 685–690, Hyderabad, India.

William Lewis and Fei Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Journal of Literary and Linguistic Computing (LLC)*, 25(3):303–319.

Michael Lockwood. 2016. Automated gloss mapping for inferring grammatical properties. Master's thesis, University of Washington.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530. Association for Computational Linguistics.

Douglas Wharram. 2003. *On the interpretation of (un)certain indefinites in Inuktitut and related languages*. Ph.D. thesis, University of Connecticut.

Dieter Wunderlich. 2003. Optimal case patterns: German and Icelandic compared. *New perspectives on case theory*, pages 331–367.

[8] See note 7.

Dieter Wunderlich. 2006. Towards a structural typology of verb classes. *Advances in the theory of the lexicon*, pages 58–166.

Fei Xia and William D. Lewis. 2007. Multilingual structural projection across interlinear text. In *Proc. of the Conference on Human Language Technologies (HLT/NAACL 2007)*, pages 452–459, Rochester, New York.

Fei Xia, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey, and Emily M. Bender. 2016. Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation*, 50:321–349.