Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2878–2883
Marseille, 11–16 May 2020

Automated Parsing of Interlinear Glossed Text From Page Images of Grammatical Descriptions

Erich R. Round[*‡§], **Jayden L. Macklin-Cordes**[*], **T. Mark Ellison**[†], **Sacha Beniamine**[§]
[*] School of Languages and Cultures, University of Queensland, Brisbane, Australia,
[‡] Surrey Morphology Group, University of Surrey, Guildford, UK,
[§] Max Planck Institute for the Science of Human History, Jena, Germany,
[†] Cologne Center of Language Sciences, Universität zu Köln, Germany,
e.round@uq.edu.au, j.macklincordes@uq.edu.au, tmarkellison@gmail.com, beniamine@shh.mpg.de

Abstract

Linguists seek insight from all human languages, however accessing information from most of the full store of extant global linguistic descriptions is not easy. One of the most common kinds of information that linguists have documented is vernacular sentences, as recorded in descriptive grammars. Typically these sentences are formatted as interlinear glossed text (IGT). Most descriptive grammars, however, exist only as hardcopy or scanned pdf documents. Consequently, parsing IGTs in scanned grammars is a priority, in order to significantly increase the volume of documented linguistic information that is readily accessible. Here we demonstrate fundamental viability for a technology that can assist in making a large number of linguistic data sources machine readable: the automated identification and parsing of interlinear glossed text from scanned page images. For example, we attain high median precision and recall (>0.95) in the identification of example sentences in IGT format. Our results will be of interest to those who are keen to see more of the existing documentation of human language, especially for less-resourced and endangered languages, become more readily accessible.

Keywords: "Information Extraction, Information Retrieval", Less-Resourced/Endangered Languages, Morphology, Typological Databases

1. Introduction

Linguistic typology is the subfield of linguistics which studies the design features of human language and the distribution of such features across the languages of the world. As in other disciplines, there has been a shift in linguistic typology towards the use of algorithmic techniques applied to computational data sets, which expands the scope of questions that can be asked (Bickel, 2007). However, there is a bottleneck at the stage of data acquisition. The primary sources of data for typologists are grammatical descriptions of languages. As most of these are in hardcopy, constructing datasets for computational use is a lengthy, manual process. Usually, the knowledge being formalised is confined to questions particular to the study at hand (Cooper, 2014, p. 91), so this process of manual database construction must be repeated with each new research question. A research pipeline automating the process of searching the published grammatical descriptions of the world's languages would greatly reduce the time-cost of novel dataset creation. However, such a goal requires initially a machine-readable compendium of grammars. Although the digitisation of linguistic resources continues apace, machine-readable grammars are available for only a small minority of the world's languages (Szymanski, 2012; Szymanski, 2013).

Computational linguistics can also benefit from more linguistic diversity in machine-readable resources. Even tools developed for a single language can be enhanced by training with other-language data. For example, Berg-Kirkpatrick and Klein (2010) find that training dependency parsers simultaneously on several languages — as opposed to just the target language — reduces the error rate for any individual language by around ten percent. Weighting languages by their phylogenetic relation to the target further reduces error. Application of these advantages is hindered by the lack of cross-linguistic data available: Szymanski (2012) estimates that, of approximately 7,000 languages recorded on earth (Simons and Fennig, 2018), around 99 percent should be considered "resource poor" in terms of computationally accessible materials. This result is even more striking when we consider that today's 7,000 languages are less than one percent of the languages ever spoken (Bickel, 2007, p. 245).

This paper covers the parsing of interlinear glossed text (IGT) in scanned images of hardcopy grammatical descriptions. IGT is a semi-standardised display format for linguistic data (Lehmann, 1982; Goodman et al., 2015), typically consisting of three lines: one for the vernacular language data (written either to a phonemic standard or in the regular orthography of the language), a gloss line giving a word-by-word translation and morphological breakdown of the data, and a line containing a free translation into English. See the IGT in (1), taken from example 78 in Odé (2002, p.63), a grammar of Mpur, a Non-Austronesian language of Irian Jaya, Indonesia.

(1) A-onsra a-bi jan mafun fan.
3SM-like 3SM-possess house beautiful many
'He likes his many beautiful houses.'

Prior work on the discovery and aggregation of IGT examples in a digital format includes the Online Database of Interlinear Text (ODIN) project (Lewis, 2006; Lewis and Xia, 2010), which uses a web crawler to search for and extract snippets of IGT contained in web pages. In contrast to the present work, ODIN is primarily focused on search capabilities — the ability for researchers to find relevant re-

sources (Lewis, 2006). Here, we focus on developing tools to identify, extract and tag IGT from source documents with a high degree of accuracy and precision. Also, while ODIN deals only with materials which are already online in a machine-readable format, we direct attention to the process of creating machine-readable representations of the scanned images of hardcopy documents, expanding coverage to languages where only older documentation exists. Szymanski (2012) also works directly on the extraction of machine-readable text from scanned grammatical descriptions. However, his aim is to extract bitext — vernacular sentences with English translations. IGT is one source of bitext, though not exclusively so.

2. Aim

Our goal is to make as many grammatical descriptions as possible available for computational analysis. Thus, our starting point is images of book pages, on which optical character recognition (OCR) needs to be performed. Consequently, we do not anticipate that the parsing of grammatical description documents can be fully automated, because 1. OCR is imperfect and 2. the conventions for representing grammatical descriptions are necessarily subject to adaptation by authors. For example, OCR often introduces errors into the spacing between words and morphemes vital to their alignment in IGT. The result of parsing these aligned line-fragments thus often needs substantial human intervention. For this reason, we do not focus on alignment in this paper, but rather on the identification of lines of IGT and the functions of those lines as vernacular text, gloss or free translation. We demonstrate proof-of-concept for the automated identification of IGT example sentences and the functions of lines that comprise them within OCRed grammatical texts.

3. Document pre-processing

Two kinds of pre-processing are applied to documents. Prior to OCR, the page images are manually marked up with plain-text areas and tabular areas. Tabular areas enclose text structures likely to be hard to parse: tables, rule formulae, non-canonical IGT with vernacular, gloss and/or translations sharing lines. Tabular areas are retained separately for later analysis. Using commercial software (ABBYY FineReader) the plain text areas, including canonical IGTs, are OCRed and output as HTML in order to preserve line breaks and italics, which are meaningful layout features in IGTs.

Grammatical descriptions typically declare the abbreviations they use. A standard for abbreviations in IGT is the Leipzig Glossing Rules (Comrie et al., 2008). Though the standard provides many of the abbreviations encountered in newer documents, most existing grammars predate it, particularly the hard copy ones that our method targets. Moreover, since the Leipzig standard provides abbreviations only for the most common grammatical category labels, documents that use it will often supplement it with additional, nonstandard items. Thus, for each grammar we create an individual abbreviations file that lists the abbreviations which the grammar declares. In addition, we specify layout regularities specific to documents, for example, a regular expression describing the format of example numbers such as (10b) or 11-2.

4. Document parsing

We parse the grammatical description document in six steps. 1. The initial HTML is transformed into an XML representation in which we preserve aspects of text layout that are potentially meaningful, such as line breaks and italics. 2. Minor annotations are made. For example, vernacular language words in the main text are identified. 3. Interlinear text is located. 4. Interlinear text is parsed. 5. Content tags are added to the XML and purely layout-related tags are removed. For example, we mark up sentences and then remove line break tags. 6. Derived outputs and reports are created. Figure 1 provides an example of the XML representation of an IGT for the example shown in 1.

4.1. Representing relevant layout

In linguistic documents, layout may be used in addition to literal content, in order to convey information about document structure, or metadata pertaining to certain pieces of text, much as indentation and italicization are used meaningfully in dictionary entries (Maxwell and Bills, 2017). In IGTs, line breaks conventionally separate vernacular from interlinear glossing from translations, while italicization conventionally marks words in vernacular languages, or English words being cited metalinguistically. Because they are often meaningful, these two typographical features are maintained in the transformation from OCR-output HTML into XML, while other layout information is discarded. We have found boldface to be largely unhelpful, since it is poorly recognised during OCR and does not serve any standard meaningful purpose, unlike italics. We do however preserve page break information, as this assists in referencing the page location of IGTs with respect to the printed original.

4.2. Minor annotations

We mark up instances of abbreviations recorded in the abbreviations parameter file, which we manually compile from the source's abbreviations list. Likely vernacular words are identified as non-English words in italics. Apparent linguistic examples in body text — likely vernacular words followed immediately by quoted text — are also marked up.

4.3. Heuristic identification of line function

To identify the functions of lines in the document, we automatically annotate all lines in plain text OCR areas with a set of features. Features include the *presence* of: abbreviations; italicized vernacular words; clause breaks (i.e., comma or semi-colon); line-final punctuation; an example number; an initial example number (e.g. 2 or 2a, but not 2b); or quotes around the majority of the text on the line, also *counts* of: plain spaces; non-breaking spaces (which is how the OCR typically renders spaces used for padding); morpheme juncture symbols; words; English words, and *edge markers* indicating whether the line is at the start or end of a plain text area.

```
<x x_number="78" html_id="seq_793+seq_794+seq_795" id="939">
      <xfg>
        <xgg xgg_id="61">
          <xv line_id="490" has_vernacular="TRUE"
               has_multiple_vernacular="TRUE" word_count="5">
            <xw>A-onsra</xw>
            <xw>a-bi</xw>
            <xw>jan</xw>
            <xw>mafun</xw>
            <xw>fan.</xw>
          </xv>
          <xg line_id="491" word_count="5">
            <xw>3SM-like</xw>
            <xw>3SM-possess</xw>
            <xw>house</xw>
            <xw>beautiful</xw>
            <xw>many</xw>
          </xg>
        </xgg>
        <xf line_id="492">He likes his many beautiful houses.</xf>
      </xfg>
</x>
```

Figure 1: Example of XML output (Odé, 2002, p. 64, ex 78).

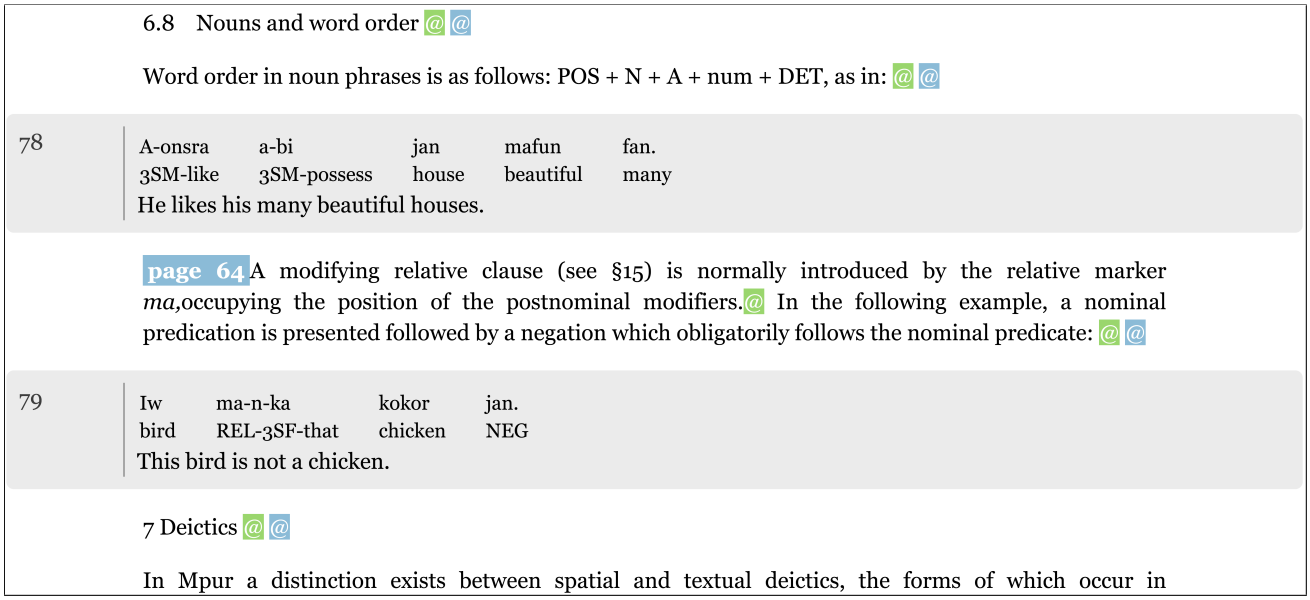
We use these features to identify four classes of lines involved in IGTs: (u) unparsed vernacular text, which contains words in the vernacular language that are not parsed into morphemes, (v) parsed vernacular text, (g) gloss, and (f) free translation. A single IGT has a hierarchical structure. A gloss line forms what we call a *gloss group* (G) with the vernacular line or lines above it: a gloss group has the form ug, vg or uvg. A free translation forms what we call a *free translation group* (F) with one or more gloss groups above it: a free translation group thus has the structure $(G^{+}f)$. A single example will contain one or more free translation groups, only the first of which is numbered, and a multi-part example may contain a set of numbered subparts, with numbering such as (11a), (11b), (11c), which also often appear in an abbreviated format such as (11a), (b), (c). For this reason, we track whether the first u or v line of any free translation group has a number, and if it appears to be simple, such as (11), or complex, such as (11b), or a part of a complex number such as (b).

This model defines the most general structure of an IGT. For each document, a manually created document parameter file permits the contribution of further constraints. For example, setting "expect_parsed_vernacular_lines" to FALSE eliminates the possibility of a gloss group having the structure vg or uvg, leaving only ug, and we specify a regular expression that describes the expected format of examples' numbers. In order to infer the functions of individual lines in the document we use a combination of line-internal cues and contextual constraints given by the model. For example, the presence of many hyphens is a line-internal cue which likely indicates a morphemically parsed line, while tht fact that free translation groups should end with f is a contextual cue given by the model. In the current implementation, these bases of inference yield results which while imperfect, are nevertheless effective, see *Evaluation*.

4.4. Analysis of typographically parsed gloss text

In IGT, vernacular and gloss lines are parsed by linguists into space-delimited words, with the words often further divided into morphemes. Sometimes, vernacular words are presented without morpheme divisions on one line and with them on the next. We parse matching vernacular and gloss lines into words, and where possible into morphemes, and check that within corresponding lines, the number of constituents matches. Apparent mismatches can appear when a single vernacular morpheme is glossed by two English words, such as 'go out'. Such mismatches lead to gloss lines whose word count is higher than the corresponding vernacular line. We identify such mismatches and attempt to repair them by comparing potential phrases to WordNet's multi-word entries (Fellbaum, 1998). OCR can introduce spaces next to hyphens and numerals. Again, this causes gloss and vernacular lines to mismatch in word count, and we attempt to repair them. OCRing can introduce spaces within morphemes and delete spaces between words. Automated detection of these errors is less straightforward, and a task for future work.

The result of these procedures is a set of example sentences structured as IGT as defined by the pattern described above.



6.8 Nouns and word order @ @

Word order in noun phrases is as follows: POS + N + A + num + DET, as in: @ @

78

A-onsra	a-bi	jan	mafun	fan.
3SM-like	3SM-possess	house	beautiful	many

He likes his many beautiful houses.

page 64 A modifying relative clause (see §15) is normally introduced by the relative marker *ma,*occupying the position of the postnominal modifiers.@ In the following example, a nominal predication is presented followed by a negation which obligatorily follows the nominal predicate: @ @

79

Iw	ma-n-ka	kokor	jan.
bird	REL-3SF-that	chicken	NEG

This bird is not a chicken.

7 Deictics @ @

In Mpur a distinction exists between spatial and textual deictics, the forms of which occur in

Figure 2: Example of HTML user-friendly rendering, source: (Odé, 2002, p. 64, ex 78)

4.5. Content representation

Once an IGT has been detected and parsed, we no longer require information contained in lines and linebreaks. We remove line nodes, and divide paragraphs into sentence nodes. Because we envisage a post-processing stage involving manual human curation, we ensure that all nodes in the XML document carry a pointer to the original HTML node, from which their contents are derived.

4.6. Derived outputs and reports

OCR is imperfect, and linguists format IGT in diverse ways. Consequently, we expect that our automated processes will need to be corrected by human curation. To facilitate this, we apply XSLT to produce a derived HTML document, that enables easy, interactive inspection of our parsed XML structure. This includes visual mark-up of example sentences, vernacular words and other annotations, as well as enabling checking of identifiers corresponding to parts of the text, that track XML nodes and original HTML nodes (revealed by hovering over coloured @ symbols). An example of this user-friendly rendering is shown in Figure 2.

In most grammatical description documents, IGT example sentences are numbered sequentially. We produce a summary report of all example sentences found, noting whether their numbering is sequential as one would expect, or departs from sequence, which likely indicates a nearby failure to parse one or more examples.

5. Evaluation

To date, we have parsed 90 grammatical description documents. Here we evaluate the IGTs identified in a sample of 16 (Table 1), containing a total of 6,671 example sentences that ideally should be parsed as IGTs. Evaluation was performed by manually examining examples sentences in the user-friendly rendition of the parsed grammar, as in Figure 2, whose layout features aid rapid visual inspection.

Language name	ISO 639-3	Document
Abau	aau	Lock (2011)
Bunuba	bck	Rumsey (2000)
Darkinyung	xda	Jones (2008)
Dhanggati	dyn	Lissarrague (2007)
Diyari	dif	Austin (1981)
Garrwa	wrk	Mushin (2012)
Giimbiyu	zme	Campbell (2006)
Gudanji	nji	Aguas (1968)
Kukatj	ggd	Breen (1992)
Mara	mec	Heath (1981)
Mende	sim	Hoel et al. (1994)
Mpur	akc	Odé (2002)
Murrinh-Patha	mwf	Mansfield (2014)
Nyulnyul	nyv	McGregor (1996)
Suboo	woi	Han (2015)

Table 1: Documents in the evaluation set

Table 2 reports precision and recall at the granularity of whole example sentences, asking how often the parser created an IGT node for example sentences in the source document, and how often the nodes it created corresponded to actual example sentences. Table 3 reports on the content of the IGT nodes: whether they exactly contained the example sentence or alternatively, missed some text (underparsed) or contained extraneous text (overparsed).

16 documents	Precision	Recall
Median	0.98	0.99
Minimum	0.86	0.74
Maximum	1.00	1.00

Table 2: Inferred IGT instances, versus original

Our median and best case results indicate that even at this initial stage, IGT can be identified with good reliability using only simple feature-based heuristics and a model of IGT

16 documents	Underparsed	Overparsed
Median	2%	2%
Minimum	0%	0%
Maximum	50%	15%

Table 3: Percentage of erroneously parsed IGTs

structure. However, worst case results even in a small sample indicate that if source documents follow unusual layout conventions, performance can deteriorate rapidly. For example, the only source document for which underparsing exceeded 0.25 (Heath, 1981) contained IGT mainly in the form of texts, not example sentences. Since our features encode expectations for example sentences, which have short free translations, underparsing leapt in this instance to 0.50, because the texts' free translations were routinely long. Likewise, recall dropped below 0.95 in only one case (Mushin, 2012), to 0.74, because IGT in the texts section of that document used a different numbering system from the preceding chapters, and we had not accounted for that in the parameter file.

Other recurring causes of underparsing include cases where free translations span two lines but only one is interpreted as part of the IGT, and cases where the first of two vernacular lines is not interpreted as part of the IGT. In future improvements, the first error could be handled by elaborating our IGT model, replacing its single free translation type f with nonfinal and final subtypes. The second could be improved by applying weighted expectations about optional IGT structures, including expectations learned from other IGTs in the document. The primary cause of overparsing is lines of regular text which happen to resemble IGT lines in multiple respects, such as by being short, containing high proportions of italicized text, or vernacular text, or initial strings that resemble example numbers, and which are adjacent to true IGT lines. Our algorithm appears to handle them haphazardly, sometimes placing them incorrectly inside adjacent IGTs. Improvements would result from a better understanding and handling of such edge cases.

6. Discussion and conclusion

The high quality of median results, the simplicity of our current processes, and the nature of our worst results all indicate that the automated detection of IGT from OCRed page images is feasible already and that further improvements can be expected.

At its broadest, the goal of our project is the extraction of machine-readable linguistic information from heritage grammars available only as scanned (or scannable) documents. One of the most accessible kinds of linguistic information is the forms which are cited in the grammar. As many vernacular sentences occur within IGTs, these are important components of grammars to parse.

In this paper, we have shown that IGTs can be successfully identified using a simple featural analysis of lines matched with a customisable template for IGTs. In future work, we envisage replacing these categorical features with probabilistic constraints to allow co-restriction of parses, and Bayesian evaluation of confidence in positing IGTs.

We plan to extract more detailed information from identified IGTs. Frequently, subparts of the vernacular and gloss lines are spatially aligned to show coreference, at the word or even morpheme level. Using an OCR engine that provides more information about positioning of identified words, such as Google's OCR API or Tesseract, and an OCR output format such as hOCR (Breuel, 2007) instead of HTML, we hope to construct systems which can identify aligned chunks within lines.

Improvements in performance should be possible through automated correction of OCR typographical errors (Hammarström et al., 2017), for example through the use of contextual expectation to correct mis-read numerals and abbreviations, which play an important role in identifying IGT lines.

A limitation of our current method is that it assumes a canonical layout of IGT, for example it does not handle free translations placed on the same line as glossing, though this variant is not uncommon in grammars. Nor does it handle more complex IGT structures with additional line types including references to primary sources and associated media files; metadata about speakers and utterance context; distinct underlying and surface phonological parses; and additional layers of morphological analysis (Round, 2013; Round, 2015). An extension of our method to cover these cases too would be valuable.

In conclusion, we argue that we have achieved in this paper one significant milestone in semi-automatically analysing heritage language grammars — the automated identification of IGTs.

Code is available freely under the GPU v3 licence on a github repository[1], and archived with DOI 10.5281/zenodo.3550760.

7. Acknowledgements

The authors gratefully acknowledge support from the Australian Research Council: via grant DE150101024 to E.R., and a Centre of Excellence for the Dynamics of Language Transdisciplinary and Innovation grant TIG0116JMC to J.M-C., E.R. and M.E. J.M-C. is supported by an Australian Government Research Training Program scholarship.

8. Bibliographical References

Aguas, E. F. (1968). Gudanji. In Estrella F. Aguas et al., editors, *Papers in Australian Linguistics No. 3*, number 14 in Pacific Linguistics Series A, pages 1–20. Pacific Linguistics, Canberra.

Austin, P. K. (1981). *A Grammar of Diyari, South Australia*. Number 32 in Cambridge Studies in Linguistics. Cambridge University Press, Cambridge; New York.

Berg-Kirkpatrick, T. and Klein, D. (2010). Phylogenetic grammar induction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1288–1297.

Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, 11(1):239–251.

[1] https://github.com/erichround/LREC_IGT/

Breen, G. (1992). Some problems in Kukatj phonology. *Australian Journal of Linguistics*, 12(1):1–43.

Breuel, T. M. (2007). The hOCR microformat for OCR workflow and results. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 1063–1067. IEEE.

Campbell, L. (2006). *A Sketch Grammar of Urningangk, Erre and Mengerrdji: the Giimbiyu languages of Western Arnhem Land*. HonoursThesis, University of Melbourne, Melbourne.

Comrie, B., Haspelmath, M., and Bickel, B. (2008). The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.

Cooper, D. (2014). Data warehouse, bronze, gold, STEC, software. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 91–99.

Goodman, M. W., Crowgey, J., Xia, F., and Bender, E. M. (2015). Xigt: extensible interlinear glossed text for natural language processing. *Language Resources and Evaluation*, 49(2):455–485.

Hammarström, H., Virk, S. M., and Forsberg, M. (2017). Poor man's OCR post-correction: Unsupervised recognition of variant spelling applied to a multilingual document collection. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 71–75.

Han, T. J. (2015). *A sketch grammar of Suboo*. PhD Thesis, Nanyang Technological University.

Heath, J. (1981). *Basic materials in Mara: grammar, texts and dictionary*. Number 60 in Pacific Linguistics Series C. Pacific Linguistics, Canberra.

Hoel, H. M., Ikäheimonen, T., and Nozawa, M. (1994). *Mende Grammar Essentials*. Unpublished Typescript, The Summer Institute of Linguistics, Ukarumpa, Papua New Guinea.

Jones, C. (2008). *Darkinyung grammar and dictionary : revitalising a language from historical sources*. Muurrbay Aboriginal Language and Culture Co-operative, Nambucca Heads, NSW, Australia.

Lehmann, C. (1982). Directions for interlinear morphemic translations. *Folia Linguistica*, 16(1):199–224.

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

Lewis, W. D. (2006). ODIN: A model for adapting and enriching legacy infrastructure. In *2006 Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*, pages 137–137.

Lissarrague, A. (2007). *Dhanggati grammar and dictionary with Dhanggati stories*. Muurrbay Aboriginal Language and Culture Co-Operative, Nambucca Heads, NSW, Australia.

Lock, A. (2011). *Abau grammar*, volume 57 of *Data Papers on Papua New Guinea Languages*. SIL-PNG Academic Publications, Papua New Guinea.

Mansfield, J. (2014). *Polysynthetic Sociolinguistics: The Language and Culture of Murrinh Patha Youth*. PhD Thesis, The Australian National University, Canberra.

Maxwell, M. and Bills, A. (2017). Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91.

McGregor, W. (1996). *Nyulnyul*, volume 88 of *Languages of the World/Materials*. Lincom Europa, München.

Mushin, I. (2012). *A grammar of (Western) Garrwa*. De Gruyter Mouton, Berlin, Boston.

Odé, C. (2002). A sketch of Mpur. In Ger P. Reesink, editor, *Languages of the Eastern Bird's Head*, volume 524 of *Pacific Linguistics*, pages 45–107. Research School of Pacific and Asian Studies, Australian National University, Canberra.

Round, E. R. (2013). *Kayardild Morphology and Syntax*. Oxford University Press, Oxford.

Round, E. R. (2015). Rhizomorphomes, meromorphomes, and metamorphomes. In Greville G. Corbett, et al., editors, *Understanding and measuring morphological complexity.*, pages 29–52. Oxford University Press, Oxford.

Rumsey, A. (2000). Bunuba. In R. M. W. Dixon et al., editors, *Handbook of Australian languages, vol. 5*, pages 34–152. Oxford University Press.

Szymanski, T. D. (2012). *Morphological Inference from Bitext for Resource-Poor Languages*. Phd thesis, University of Michigan.

Szymanski, T. D. (2013). Automatic extraction of linguistic data from digitized documents. In *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society*, volume 39, pages 273–286.

9. Language Resource References

Fellbaum, Christiane. (1998). *WordNet: An Electronic Lexical Database*. MIT press.

Simons, Gary F. and Fennig, Charles D. (2018). *Ethnologue: Languages of the world*. SIL International, 21st.