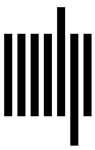
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

3 The Scope of Linguistic Data

Jeff Good

1 The range of linguistic data

There is tremendous diversity in the kinds of data used in the study of language, which reflects the unusual position of linguistics as a discipline where scholars frequently adopt methods associated with the humanities, social sciences, cognitive sciences, and computer science, among other areas. The goal of this chapter is to discuss an illustrative range of linguistic data types within a general classificatory framework that can, in principle, be extended to kinds of data beyond those directly considered here.[1] This framework is offered in the spirit of starting a broader discussion of how linguists can classify the data on which their scholarship is based, and that, in turn, should allow for more informed consideration of issues surrounding data management. This survey is necessarily incomplete because covering all the kinds of data used in linguistic research would require more space than is available. Due to the nature of my own expertise, this survey is somewhat biased toward linguistic data associated with language documentation and description, though attempts have been made to highlight important kinds of data throughout the discipline, and this survey can be usefully complemented by consideration of the kinds of data described in many of the other chapters in this volume.

The foundation of the study of language is data derived from observable linguistic behavior, broadly construed here to include both naturalistic and elicited data (section 2). Data of this kind can be subjected to a wide range of analyses, and these analyses can produce new kinds of linguistic data that become the subject of further analysis. The resulting analyses can take the form of diverse kinds of annotations, representations of syntagmatic and paradigmatic structure, and representations of lexical information, among other possibilities (section 3). These can then form the basis of generalizations about specific languages or language in general that serve as data for studies looking at topics such as worldwide patterns of language variation or cognitive linguistic universals (section 4).

Alongside data from specific languages and about language more generally, there are a range of data types of importance to linguistic investigation that go beyond language itself. The most significant of these is data about individuals, which, when aligned with data on their language use, are central to subfields such as sociolinguistics and anthropological linguistics (section 5). Increasing attention has also been placed on metadata in linguistics, in particular to support the archiving and discovery of language resources (section 6).

2 Data from observable linguistic behavior

2.1 Observable linguistic behavior, broadly construed

A core tenet of linguistics is its adoption of a descriptive, rather than prescriptive, approach to the analysis of language. This entails the collection of linguistic data that are directly observable, though the field is divided on what kinds of observable data can be considered valid as the basis of linguistic analysis, roughly along so-called functionalist and formalists lines (see, e.g., Newmeyer 1998). In broad terms, more functionally oriented linguists emphasize the importance of naturalistic instances of language in use as foundational data for linguistic investigation, often under the heading of usage-based approaches (see, e.g., Langacker 1987:46; Diessel 2017). By contrast, more formally oriented linguists see it as appropriate to rely on constructed examples of language that can serve as prompts to collect grammaticality judgments from users of a language. Schütze ([1996] 2016) provides relevant critical consideration of this kind of data.

Here, these two classes of data are covered under the broad category of data derived from observable linguistic behavior to contrast them with data that are based on the analysis of such observations. Data from language use are further discussed in section 2.2, which focuses on language documentation, and section 2.3, on textual corpora. In section 2.4, more specialized kinds of data based on observable behavior are considered, including grammaticality judgments and information collected via technical instruments.

2.2 Documentary linguistic data

While naturalistic data of language use can be used to support almost any kind of linguistic work, they are central to one subfield in particular, documentary linguistics, which is based around "the creation, annotation, preservation, and dissemination of transparent records of a language" (Woodbury 2011:159), in particular in contexts of language endangerment. As such, the documentary linguistics literature contains fairly extensive consideration of different kinds of naturalistic data that can be collected and the methods that can be used to facilitate their collection (see also Cox, chapter 22, this volume; Daniels & Daniels, chapter 26, this volume).

Himmelmann (1998:180), for instance, widely cited as the first work to contrast documentary linguistics with other areas of the field, specifically discusses the notion of a "systematics of communicative events" to help those engaged in the documentation of a language to ensure that the data they collect are not merely naturalistic but also representative of the actual linguistic practices of a community. He further suggests that a parameter of "spontaneity" (178) can help structure data collection to produce a more accurate record of a language. The issue of representativeness is a broad one given that it necessarily encompasses not only different genres but also diversity among members of a linguistic community across dimensions such as age, gender, and other culturally significant social groupings (see, e.g., Childs, Good, & Mitchell 2014 for general discussion).

Alongside considerations of what kinds of events to record, a central concern of documentary linguistics has been the mechanics of data collection, as evidenced by work on data management (e.g., Austin 2006; Good 2011; Thieberger & Berez 2012) or audio and video recording techniques (e.g., Margetts & Margetts 2012). An important recent development in documentary linguistics has been increased attention on the collection of video data (see, e.g., Dimmendaal 2010; Seyfeddinipur 2012; Seyfeddinipur & Rau 2020). Obviously, for sign languages, proper documentation is inconceivable without video recording (see Schembri 2010:112–116). For spoken languages, audio recording can produce records that can support many kinds of linguistic analysis effectively. However, to the extent that interactional communication, even when primarily being accomplished via speech, typically involves a visual component (e.g., via gesture or gaze), many researchers in documentary linguistics have determined that the visual context of a speech event constitutes a valuable kind of data for linguistic analysis, even if it is only arguably "language" data.

While work that situates itself specifically within documentary linguistics tends to focus on endangered varieties, the kinds of data that it focuses on can be collected for any language and similar approaches have been adopted in other subfields, as evidenced for instance by the data sets assembled as part of the TalkBank project (MacWhinney 2007) or speech data used as the basis for sociolinguistic investigation (Kendall 2008, 2011; see also Sonderegger et al., chapter 15, this volume; Kendall & Farrington, chapter 14, this volume; Fridland & Kendall, chapter 18, this volume). Data of this kind can be considered to be documentary in nature, even if language documentation as a term is typically applied to endangered language contexts.

The notion of documentary linguistic data rests on the idea that it is possible to record linguistic events that can be considered "naturalistic" despite the fact that the act of recording them is not a naturalistic part of the event. This issue has been frequently referred to under the heading of the *observer's paradox* (Labov 1972:209; see also Birch 2014:32–34). However, the range of ways that the act of observation may alter patterns of language use across recording contexts and cultures does not yet appear to have been the subject of general investigation.

2.3 Textual corpora

Another significant category of data derived from observable linguistic behavior is textual data. While such data can, in principle, be drawn from any written text for analysis, they most clearly become linguistic data when assembled into a corpus of some kind (see McCarthy & O'Keeffe 2010 for a historical overview linguistic corpora). The term *corpus* can be used broadly to cover both text corpora and audiovisual corpora (of the sort discussed in section 2.2), but the subfield of corpus linguistics is

most strongly oriented toward textual analysis (see, e.g., Bonelli 2010:18–19; Gries & Berez 2017:380–381). Moreover, corpus linguistics is typically based on the analysis of textual data when that data can be considered primary data rather than annotations on primary data, as would be the case, for instance, of a transcription of an audio recording (see section 3.2 for consideration of annotation), though, once a transcription exists, the same analytical methods can be applied to the resulting textual record.[2]

The existence of textual corpora highlights the fact that, in societies characterized by widespread literacy, textual data can be significant sources of observable linguistic data. Data from journalistic sources, in particular, have played an important role in the development of large-scale corpora (Bonelli 2010:16), both because of their availability and because they represents a linguistic genre that is naturally text based. Beal, Corrigan, and Moisl (2007:1–2) make a distinction between conventional and unconventional corpora. The former focus on varieties that are associated with standardized writing systems. Examples include the *British National Corpus* (BNC Consortium 2007) (see also Gries, chapter 38, this volume) or the Corpus of Contemporary American English (Davies 2008–). While there is significant space for variation within conventional corpora, there is much less variation than in unconventional corpora, which are based on more heterogeneous input data sources. The corpora on creole languages described by Sebba and Dray (2007) provide an example of the potential complications. For instance, creole language text may be interspersed with text from a standardized language, such as English, in a novel raising questions of just what should be included in a corpus from such a source (189).

Well-developed corpora are not limited to collections of texts themselves but can also contain annotations of the texts and ancillary resources such as lexicons to assist in their interpretation (see section 3). Strassel and Tracey (2016) discuss these kinds of corpora, describing them using the term "language pack." In contrast to this is the increasing use of more ad hoc corpora derived from texts made available online, in particular via services that regularly aggregate new instances of naturally generated text, such as Twitter (see, e.g., Grieve, Nini, & Guo 2017; Scannell, chapter 41, this volume).

Text corpora highlight the dual role of textual data in linguistics in that they can sometimes serve as primary data, as is the case for typical corpora, while in other instances, the textual representations are seen as secondary representations of some other kind of primary data, as is typically the case in language documentation (see section 2.2). They also highlight the role of curation in the creation of linguistic data. Some text corpora, such as the textual portion of the British National Corpus (British National Corpus 2007), are highly curated so that the resulting corpus can be considered representative of a certain set of linguistic varieties. A resource such as News on the Web (Davies 2013), which is based on a selection of online news resources and continuously updated to reflect newly available content, reflects a more passive curatorial approach.

2.4 Specialized data from observable behavior

The kinds of data discussed to this point can be broadly described as "naturalistic" insofar as they are intended to be reflective of actual language use. By contrast, there is one very prominent kind of linguistic data that comes from observable linguistic behavior, but of a highly specialized nature. This involves language user judgments of the acceptability of a given expression. The most well-known class of judgments of this kind are so-called grammaticality judgments (see, e.g., Schütze [1996] 2016 for critical consideration of this kind of data and Sprouse 2013 for a bibliographic overview).

As discussed by Abrusán (2019), there are various possible reasons why a given expression can be considered unacceptable, and they could be primarily syntactic (or morphosyntactic), semantic, or pragmatic, with the details dependent on the expression in question as well as the context in which the expression is interpreted (see also McCawley 1998:5–6). For certain theoretical approaches to linguistics, in particular generative approaches, language data consisting of sets of sentences associated with judgments of their acceptability play a central role in the analytical process.[3] In terms of presentation, sentences deemed to be inappropriate are generally annotated (see section 3.2) with "stigmata" (see McCawley 1998:3) classifying the nature of their unacceptability. The asterisk (*) is the best known of these stigmata and is typically used as a marker of syntactic ungrammaticality.

While not as theoretically prominent, language user judgments are also used to study non-syntactic domains of grammar, with a well-known example involving judgments as to whether specific sound sequences are considered to be possible words in a language even if those sequences happen not to be associated with a

given word. A frequently cited example for English is an opposition between the non-words *blick* and *bnick*. The former is generally judged to consist of a sequence of sounds that could be a word in English, while the latter is judged to not be a possible word due to its initial *bn* sequence (see, e.g., Cohn 2001:180).

These data are clearly of observable linguistic behavior, though of a highly specialized kind of behavior specifically designed to facilitate linguistic research. Rather than proposing a categorical distinction between data of this kind and naturalistic data, it is probably better to see these as different ends of a continuum of control in data collection (see Birch 2014:27–29), with data gathered on acceptability judgments being at the highly controlled end of the continuum. While different theoretical approaches may weigh this data more or less heavily with respect to linguistic analysis, this does not change the status of this kind of data as emanating from language "use," albeit of a very atypical kind.

Additional kinds of specialized data on language use involve the collection of fine-grained aspects of linguistic production or perception via instrumental means. An early instance of this kind of data is the palatogram (Ladefoged 1957), which is a record (e.g., in the form of a photograph) of where a substance that has been placed on the palate has been removed due to the movement of the tongue. Another early instance of this kind of data are spectrograms (Koenig, Dunn, & Lacy 1946), a standard part of the tool kit of phonetic analysis, now widely used even by non-phoneticians due to the availability of tools such as Praat (Boersma & Weenink 2019), which make them easy to generate. There is no single catalog of instrumental data that are used in linguistics, and they can clearly take on quite diverse forms. Phillips and Wagers (2007:747), for instance, list various kinds of instrumental data used in psycholinguistic studies such as eye-tracking in self-paced reading tasks and event-related potentials, which can measure electrical brain activity in response to a linguistic stimulus and are based on electroencephalographic measurements (see Kaan 2007; Beres 2017). A similar kind of instrumental data that is increasingly being used is functional magnetic resonance imaging, also to measure brain activity (see, e.g., Willems & van Gerven 2018).

Specialized kinds of data on language use can be contrasted with data of the kind associated with language documentation (see section 2.2) or data from corpora (see section 2.3) by the fact that they are generally collected with a very specific analytical goal in mind, whether this is the formal analysis of a syntactic pattern, modeling the articulation of a particular sound, or understanding how a given linguistic construction is processed. By contrast, documentary and corpus data are generally likely to be usable to support a wide range of investigations across more than one linguistic subfield, though data of such kinds could also be collected to serve a fairly narrow purpose depending on the research practices adopted.

3 Analytical structures applied to data of language use

3.1 Building analyses onto observable data

Most linguistic investigation is not based directly on representations of observable linguistic behavior, but, rather, on data derived from analyses of these observations. This is probably seen most directly in the field's reliance on written representations of linguistic data, whether in the form of transcription systems or orthographies that are used to approximate spoken or signed forms. These represent one kind of possible annotation that can be made on a linguistic data source; annotation in general will be discussed in section 3.2. Linguistic data can also be arranged in ways that facilitate abstract analysis, and two important kinds of structural analysis, across the syntagmatic and paradigmatic dimensions, are considered in section 3.3. The special case of lexical data is considered in section 3.4.

The topics in this section are somewhat heterogeneous in nature. Annotation, for instance, refers to a way of encoding analyses rather than representing any specific kind of analysis, and annotation can, in principle, be used to encode syntactic, paradigmatic, or lexical analyses, for instance. These topics are grouped together as part of a consideration of the kinds of linguistic data that are generated via the analysis of data based on observable linguistic behavior.

3.2 Linguistic annotation

Linguistic annotation is a kind of linguistic data that "involves the association of descriptive or analytical notations" with other kinds of language data (Ide 2017:2). Annotation can either be made directly on "raw" data (i.e., unannotated language data) (see, e.g., Schultze-Berndt 2006:215; Himmelmann 2012:188) or on other annotations.

To make the discussion more concrete, consider the example in (1) from the language Yeri [glottocode: yapu1240],[4] drawn from Wilson (2017:29). This example represents one of the more commonly encountered kinds of annotated data seen in linguistic analysis, interlinear glossed text.

(1) *hem* *ta* *m-y-aya* *maŋa-Ø?*
1SG FUT 1SG-2-give.R what-SG.R
"What will I give you (sg. or pl.)?" (120517–001:185.991) RNS, JS

Interlinear glossed text is a data format geared toward the presentation of linguistic data from languages other than the language that is being used to describe the data (e.g., English, French, or Russian). It provides a visually compact means of providing translational equivalents under each word of the language being analyzed, typically referred to as *glosses*. It can potentially include an indication of morpheme boundaries (signified by hyphens in (1)) and the use of abbreviations for grammatical terms in the glosses, often presented using distinctive typography (e.g., small capital letters in (1)).[5] In addition to presenting word-by-word glosses, interlinear glossed text is also typically associated with a free translation of the entire linguistic fragment being analyzed, as is found in the third line of (1). Depending on the presentational needs of a given work, interlinear glossed text may include additional lines, for instance a line including a representation of the relevant linguistic fragment in a distinct script from the one being used to present the analysis (e.g., a Cyrillic orthographic representation in addition to a Roman transliteration) or one line representing the linguistic fragment with morpheme boundaries and another without morpheme boundaries. Further discussion of interlinear glossed text in the context of considerations of annotation can be found in Bow, Hughes, and Bird (2003), Palmer and Erk (2007), and Goodman et al. (2015). The Leipzig Glossing Rules (Bickel, Comrie, & Haspelmath 2008) have emerged as a de facto standard for the presentation of interlinear glossed text.

In addition to presenting information needed to understand the linguistic structure of the Yeri sentence, the example in (1) also includes information identifying its source in its last line. Specifically, it is drawn from a recording with identifier 120517–001, and it begins 185.991 seconds into the recording. The abbreviation RNS found after this provides information on the genre of the collected text, which stands for recorded natural speech, and the sequence JS provides the initials of the speaker, John Sirio (Wilson 2017:28).

The example in (1) can be seen as providing multiple layers of annotation. In its first line, two kinds of annotation are provided: a written representation of the reported utterance paired with a basic morphological analysis, indicated with hyphens and, in one case, a zero-morpheme treated as part of the language's inflectional system. The second line also provides some morphological analysis in its association of each morpheme in the first line with a simple English translation or a morphosyntactic category. The final line contains four discrete annotations: a free translation and three pieces of metadata (see section 6) about the source of the example, the nature of the event from which the data was collected, and the speaker of the fragment. This example, thus, provides some sense of the diversity of possible linguistic annotations.

Interlinear glossed text should be primarily understood as a presentation format for encoding specific kinds of annotations insofar as it is optimized for visual interpretation on a page rather than encoding the data in a machine-readable or archival format (see, e.g., Bird & Simons 2003:565; Simons 2006; Good 2011:227–228; Thieberger & Berez 2012:94–96 for relevant discussions). It can be considered a specific instantiation of a general data class of morphologically analyzed texts, which can take on different forms. It has an especially compact presentation and can be compared, for instance, with the presentation of analyzed texts found in Boas (1911), which contains both interlinear translations at the level of the word, though more along the lines of a free translation rather than a gloss, along with footnotes for each word in the text providing further morphological notes, which are sometimes quite detailed.

Linguistic annotation provides an open-ended means of generating linguistic data on the basis of other linguistic data. It has taken on increasing importance as computational methods play a more central role in linguistic research because the ability of a machine to process linguistic data often relies on the presence of well-structured annotations on that data, and work on the digital encoding of linguistic annotations is where the properties of annotations have been most fully explored (see, e.g., Romary & Witt 2014; Ide 2017). An important distinction to be made in this regard is between the conceptual model underlying a system of annotation and the concrete

format that is used to express the content of the model (see Ide et al. 2017; Pustejovksy, Bunt, & Zaenen 2017). (See also Han, chapter 6, this volume, for consideration of issues connected to data transformation, which is often relevant when creating and processing annotations.)

Returning to the example in (1), the underlying conceptual model on which the interlinear glossing is based is largely implicit and also somewhat complex. It relies on the notions of word and morpheme, both of which are central to the presentation of the alignment of the elements in the first and second lines, as well as an analysis of the abstract grammatical categories that are required to understand morphosyntactic patterns in the language. The model additionally incorporates some notion of free translation. These elements are common to interlinear glossed text generally, but this particular example combines a model for linguistic analyses with a separate conceptual model for the metadata found after the free translation. As indicated, this metadata model includes information on how the annotations relate to the data being annotated, the genre of the event from which the data are drawn, and the speaker. These annotations do not constitute anything like a "complete" analysis of the data. The morphosyntactic analysis is elaborated in Wilson's (2017) descriptive grammar, and the metadata for the record are elaborated in an archival deposit (Wilson 2014).[6]

For purposes of illustration, a partial schematization of the conceptual structure underlying the presentation in (1) can be seen in figure 3.1. The recording on which the annotations are based is associated with a metadata record as well as a transcription of a specific time segment, indicated in gray. The transcription is, in turn, associated with a word-level parsing, and each word is associated with a gloss. The transcription is also associated with a free translation. Lines are used to represent annotation relations, where an element found below another element can be interpreted as an annotation on the higher element.

The schematization presented in figure 3.1 begins to represent the complexities involved in the structuring of annotations, though some only implicitly. For instance, the word-level analysis associated with the transcription represents a series of annotations that subdivide the higher-level annotation. Other annotations, such as the association of a word to a gloss represent a one-to-one relation. Both the transcription annotation and the metadata annotation are associated with the recording, but the metadata are associated with the entire recording, while the transcription is associated with a fragment of the recording that can be defined with respect to a particular time span.

In a discussion of the annotation capabilities of the ELAN multimedia annotation tool, Brugman and Russel (2004:2068) discuss a number of types of possible annotations that can be used in the analysis of multimedia data sources. These include (i) an annotation directly linked to a specific time span of a recording, (ii) a time-linked subdivision for a series of annotations that exhaustively divide an annotation linked to a time span without any gaps (e.g., a segmental annotation of a word for phonetic analysis), (iii) a subdivision without links to specific times that exhaustively divides a higher-level annotation (e.g., a word divided into morphemes when it is either difficult or unnecessary to directly link the

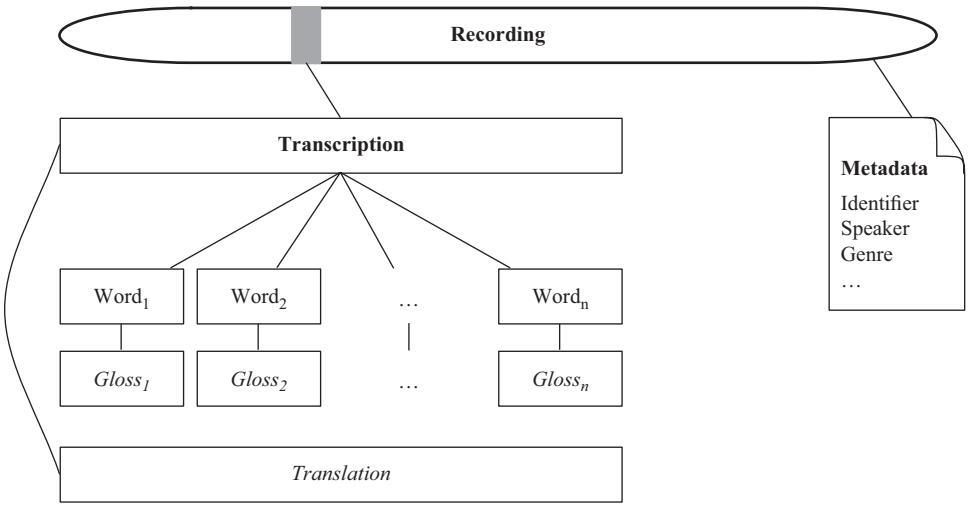


Figure 3.1
Schematization of annotation structure for interlinear glossed text.

morpheme annotations to specific time spans), and (iv) a one-to-one association where a higher-level annotation can be linked only to a single additional annotation for some kind of information (e.g., a part of speech annotation for a word). Another prominent kind of annotation in linguistics involves the association of syntactic tree structures onto textual representations of expressions to create a treebank (see, e.g., Abeillé 2003 for an overview). Detailed discussion of general issues in annotation and numerous case studies, largely from a computational perspective, can be found in Ide and Pustejovksy (2017).

In addition to the issue of the conceptual model underlying an annotation system, there is also the question of the format used to encode an annotation. A high-level distinction centers on an opposition between inline and stand-off annotation (Ide et al. 2017:79–80). The annotation system schematized in figure 3.1 presents an instance of stand-off annotation where the annotations are not included in the primary data file (in that case an audio recording) but, rather, are stored separately. Inline annotation involves embedding the annotations directly with the primary data. An example of this kind of annotation, drawn from Ide et al. (2017:80) is provided in figure 3.2. In this example, annotations are made on the linguistic fragment *Many cultural treasures are, however, not in a representative state. We have to restore them*. Unlike the example in (1), this text fragment should be considered the primary data source rather than being seen as a transcription of some underlying recording (see section 2.3).

The annotations in figure 3.2 are represented using XML (see Gippert 2006:352–361 for an accessible overview). XML is a markup language designed for (among other things) adding annotations to textual data, and this is done via a system of opening and closing tags. In figure 3.2, the opening tags are <S>, for sentence, <FACILITY>, a term being used in this context to refer to a human-made entity serving as a location, and <GROUP>, for a group of people. Closing tags are the same as opening tags except for the addition a slash before the tag name. This example partially illustrates a tagging system designed for named entity recognition, an information extraction process designed to locate sequences of text specifically referring to entities of various kinds (e.g., people, places, or organizations) (see Nadeau & Sekine 2007).

```
<S><FACILITY>Many cultural treasures</FACILITY> are,
however, not in a representative state.
<GROUP>We</GROUP> have to restore
<FACILITY>them</FACILITY>.</S>
```

Figure 3.2
An example of inline XML annotation.

Annotations both describe existing data and create new kinds of data. For instance, in (1), the written representation of the utterance in the first line can be understood simultaneously as enriching the original recording and creating new written data on the language being described that can serve as the input to further annotation and analysis, as illustrated in the conceptual model in figure 3.1. In addition, many linguistic claims that can be considered data for linguistics, especially generalizations about specific languages (see section 4) that are typically presented in the form of prose could also be reconceptualized (at least partly) as annotations.

Significant work remains on how to do this effectively, and most annotations represent relatively simple kinds of analyses. There has been work, in particular, regarding how the generalizations included in descriptive grammars could be encoded in a machine-readable form using annotations of some kind (see, e.g., Good 2004; Thieberger 2009; Bender et al. 2012; Maxwell 2012; Nordhoff 2012). However, this does not yet seem to have had a significant impact on practices in the field.

3.3 Modeling syntagmatic and paradigmatic structure

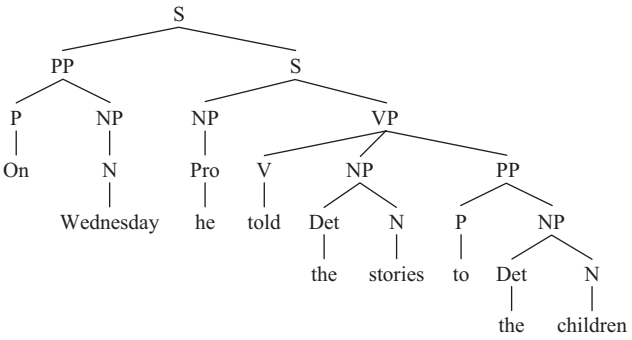
Annotations, as described in section 3.2, are a general-purpose method to associate different kinds of information with each other, and their use is not limited to linguistic data. By contrast, the abstract kinds of structural analyses described in this section are central to linguistic investigation and also create important kinds of linguistic data. These are analyses of syntagmatic structure and paradigmatic structure.

As discussed by van Marle (2000:225), syntagmatic relations hold among linguistic elements comprising some kind of linguistic constituent, while paradigmatic relations are based on a vaguer notion of "relatedness" or "connectedness" among linguistic elements within a language. The arrangements of words in a phrase, including head-dependent relationships among its subconstituents, are a well-studied kind of syntagmatic relationship, though syntagmatic relations are not restricted

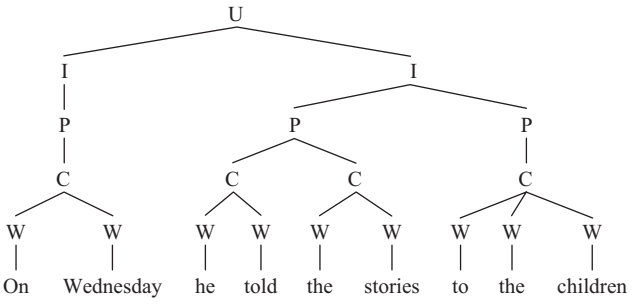
to syntactic structures but can apply to the arrangement of elements within any kind of linguistic constituent (e.g., morphological or prosodic constituents). Well-known kinds of paradigmatic relationships involve inflectional paradigms ranging from the relatively simple singular/plural opposition on English nouns to highly elaborated paradigms found in languages making use of extensive verbal or nominal morphology. However, the notion of a paradigmatic relationship is also broader than this including, for instance, the use of minimal pairs to establish phonemic contrasts in a language or the juxtaposition of two sentences with the same truth conditions (e.g., active and passive variants of a sentence) as a means to establish the syntagmatic relationship of syntactic constituency. Cataloging the entire range of possible syntagmatic and paradigmatic relationships used in linguistic analysis is outside the scope of the present chapter. However, what can be considered is the way that analyses of these relationships become encoded as linguistic data and therefore subject to further analysis.

Two representations of analyses of syntagmatic relations in the sentence *On Wednesday, he told the stories to the children*, drawn from Hayes (1990:86), are presented in (2) and (3). Following common linguistic practice, these relationships are modeled in the form of trees whose nodes are annotated for their linguistic type, thus allowing them to simultaneously encode partonymic (i.e., part-whole) and taxonomic (type-subtype) relations (see Moravcsik 2010 for further discussion). Trees can be understood as a kind of graph, in the technical mathematical use of the term as found in graph theory (see, e.g., Diestel 1997 for an introductory text). A graph, in this sense, is understood as a set of *nodes* (also termed *points* or *vertices*) and *arcs* (also termed *edges* or *lines*) connecting those nodes (see, e.g., Diestel 1997:2). Trees represent a subset of possible graphs and are constrained in a number of ways, for instance by the requirement that they have one and only one root node and that they do not allow "loops"—that is, each node can be dominated by only one other node (see McCawley 1982:91–94, 1998:46–48 for further discussion).

A standard device for representing trees in linguistic work is via tree diagrams of the sort seen in (2) and (3). The tree diagram in (2) presents a possible syntactic constituency analysis for the sentence, and the one in (3) presents a possible analysis of the sentence's prosodic constituency. The labels in (3) stand for the following prosodic constituent types: W=word, C=clitic group, P=phonological phrase, I=intonational phrase, U=utterance. In both cases, the tree diagrams represent relations among the subconstituents of a given a constituent, thus making them syntagmatic in nature.



(2)



(3)

The use of paradigmatic relationships in linguistic analysis can be illustrated by an examination of the data in table 3.1, illustrating tonal patterns for words in the Mande language Kpelle [glottocode: libe1247]. The forms are adapted from Hyman (2011:207) and based on Welmers (1962:86). The data are arranged to exemplify the range of attested surface tonal patterns in the language and to demonstrate the relatively limited number of tonal melodies found on words.

Tone-bearing units in words can surface with high (H), low (L), mid (M), or falling (F) tones. The table further includes an abstract analysis of these patterns that is facilitated by presenting them in paradigmatic opposition, namely that the surface tonal variation can be reduced to five underlying tonal patterns involving just a two-way high/low tonal opposition if one assumes various rules of tonal association, for instance that a high and low tone appearing on a single tone-bearing unit are realized as a falling tone, as well as a rule simplifying a

low-high sequence to a mid-tone (see Hyman 2011:207 for further discussion).

The paradigmatic relationships in the table 3.1 can specifically be found in the arrangement of words in the first column. They have in common that they are words found in the same language. They differ in their surface tonal patterns. For paradigmatic comparison to yield sensible results, the elements being compared must have enough in common to make it possible to provide a linguistic analysis of the source of their differences, though because of the heterogeneous nature of the notion of linguistic relatedness (see van Marle 2000:226), the range of paradigmatic comparisons that are used in linguistic analysis is quite broad. There does not appear to be extensive work surveying the kinds of paradigmatic analyses used in linguistics, though some sense of this can be found in the study of Penton et al. (2004) who propose a general model for the encoding of information found within paradigms, broadly construed to encompass "any kind of rational tabulation of words or phrases to illustrate contrasts and systematic variation" (Bird 1999:33). The paradigms that they consider include not only presentations of morphological and phonological oppositions but also sociolinguistic variants and historical cognate sets.

Table 3.1
Tone patterns in Kpelle

Word	Gloss	Surface	Underlying
pá	'come'	H	H
láá	'lie down'	HH	
ɓóá	'knife'	HH	
pílí	'jump'	HH	
kpòò	'padlock'	LL	L
tɔ̀nɔ̀	'chisel'	LL	
tòlòŋ	'dove'	LL	
kpàkì	'loom'	LL	
yɛ̂	'for you'	F	HL
kpôŋ	'door'	F	
tóà	'pygmy antelope'	HL	
kálì	'hoe'	HL	
kpōŋ	'help'	M	LH
sɛ̄ɛ̄	'sit down'	MM	
sūā	'animal'	MM	
kālī	'snake'	MM	
tɛ̄ɛ̂	'black duiker'	MF	LHL
yūɔ̌	'axe'	MF	
kōnâ	'mortar'	MF	
kpānâŋ	'village'	MF	

Unlike syntagmatic analyses, where there is a relatively standard means to model them in the form of trees, there is no standard means of representing paradigmatic oppositions. Tabular presentations, such as what is seen in table 3.1, are quite commonly employed, but, as shown by Penton et al. (2004), while tabular presentations of paradigmatic data share a presentational similarity, this masks potential complexity in the kinds of information that are presented. To pick an instance of this in table 3.1, the table appears to be structured primarily across two dimensions. The first is the vertical dimension of words with different tonal patterns, and the second is information about those words in the form of an orthographic representation, a gloss, and surface tone melody. However, there is also an implicit third dimension of information relating to a word's underlying tonal category presented both via a tonal underlying form in the fourth column of the table and via horizontal lines separating different blocks of words.

Adopting the relatively broad sense of paradigm employed by Penton et al. and applying it to the analysis of all paradigmatic oppositions provides a framework for understanding the distinction between viewing data from a synchronic and diachronic perspective in linguistic analysis. Synchronic analysis can be viewed as the analysis of data sets that are not paradigmatically opposed across the dimension of time while diachronic analysis would then be viewed as the analysis of data sets that are paradigmatically opposed across time (as well as other possible dimensions of variation). Synchronic analysis could involve the comparative analysis of linguistic varieties attested at different times (e.g., if the Latin case system were considered alongside the Finnish case system as part of a typological study of case). However, what makes a given analysis or set of data "diachronic" is that the dimension of time is considered important to the investigation.

An analytical construct of relevance in this context is the feature structure (see Sag, Wasow, & Bender 2003:50–58 for an accessible introduction; Carpenter 1992 for detailed formal consideration, and Romary & Witt 2014:183–186 for discussion in the context of linguistic annotation). These are sets of feature–value pairings that are grouped together to describe some linguistic element. Depending on the analytical approach being adopted, the value of a feature can be another feature

structure, allowing for complex, nested structures. Feature structures have been used as central analytical devices in syntactic frameworks such as head-driven phrase structure grammar (HPSG) (Sag, Wasow, & Bender 2003) and lexical-functional grammar (Bresnan 2001) and represent a flexible and powerful of way encoding both syntagmatic and paradigmatic analyses. A representation of a feature structure used to express the syntactic properties of the name *Kim* in English, drawn from Sag, Wasow, and Bender (2003:474), in HPSG is provided in figure 3.3. This feature structure is represented in the form of an attribute–value matrix where feature names are presented in capital letters and their associated values are presented to the right of the feature name. HPSG feature structures can additionally be associated with a specification of the type of linguistic object being described by the feature structure, indicated with an italicized label in the upper-right corner of an attribute value matrix in figure 3.3.

Full details on how to interpret the attribute–value matrix representation in figure 3.3 can be found in Sag, Wasow, and Bender (2003). In broad terms, it is used to present analysis of *Kim* that includes a specification of its phonological form (PHON), syntax (SYN), argument structure (ARG-ST), and semantics (SEM). The phonological representation of *Kim* is provided using English orthography. The syntactic features assigned to *Kim* are that it is of type *noun* and that it participates in a third singular agreement pattern. Because *Kim* does not take arguments

$$
\begin{bmatrix}
\textit{word} & \\
\text{PHON} & \langle\ \text{Kim}\ \rangle \\
\text{SYN} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} & \\ \text{AGR} & \textit{3sing} \end{bmatrix} \end{bmatrix} \\
\text{ARG-ST} & \langle\ \ \rangle \\
\text{SEM} & \begin{bmatrix} \text{MODE} & \text{ref} & \\ \text{INDEX} & \textit{i} & \\ \text{RESTR} & \left\langle \begin{bmatrix} \text{RELN} & \text{name} \\ \text{SIT} & \textit{t} \\ \text{NAME} & \text{Kim} \\ \text{NAMED} & \textit{i} \end{bmatrix} \right\rangle \end{bmatrix}
\end{bmatrix}
$$

Figure 3.3
Attribute–value matrix representation of a feature structure for the name *Kim*.

(which would normally be expected of verbs rather than nouns), it is listed as having an empty argument structure. The semantic properties of *Kim* indicate that it is used to refer to an entity with that name. What is of note in the present context about this complex feature structure is that it simultaneously encodes syntagmatic information—specifically, in its presentation of a phonological representation of the word and its specification that the word requires no additional arguments to be syntactically realized—and paradigmatic information—for instance, in its specification that the word is associated with the third singular agreement class. This reflects the fact that fully analyzing any syntactic constituent requires knowledge of both its syntagmatic and paradigmatic properties.

As is the case with annotations (see section 3.2), syntagmatic and paradigmatic analyses of linguistic data can themselves become data for further levels of analysis. Perhaps the most prominent case of this in linguistics is the treatment of syntactic trees as the basic units of syntactic theorizing in transformationalist approaches rather than, for instance, strings of segments (see, e.g., McCawley 1982:92). An example from the paradigmatic domain involves the study of syncretism, a phenomenon whose investigation assumes the existence of morphological paradigms and is detected by looking for patterns of formal identity in parts of those paradigms (see, e.g., Baerman, Brown, & Corbett 2005:13).

As indicated, syntagmatic and paradigmatic analyses are abstract in nature. For them to be conveyed, they must be encoded in some way. This could be done informally via prose, for instance, or via formats optimized for presentation on the printed page of the sort seen in (2) and (3) and in table 3.1. They can also be encoded in the form of annotations (see section 3.2). Annotations and structural analyses are not mutually exclusive but rather represent a distinction between a frequently used device to represent analyses (e.g., as annotations) against different conceptual kinds of analyses (e.g., syntagmatic and paradigmatic).

3.4 Lexical data

The final kind of data to be considered in this section are lexical data (see also Beier & Michael, chapter 24, this volume). While it would be logically possible to treat lexical data as a hybrid data class containing some syntagmatic and some paradigmatic information (comparable to what was seen in figure 3.3), in practice lexical data have a special place in linguistic analysis, which

is why they are treated separately here. The importance of lexical data (especially when presented in the form of dictionaries) has led to the development of the distinct field of inquiry known as lexicography (see, e.g., Durkin 2016), which is somewhat separate from the field of linguistics as whole. This can be contrasted with the study of the grammars (in the sense of descriptions of languages written by linguists). While there are works on the topic of grammaticography (see, e.g., Mosel 2006), these are quite limited and grammaticography has not developed into a separate field of inquiry in the same way lexicography has.

In part due to the widespread use of lexical resources in computational applications, lexical data have been the subject of especially extensive investigation as a data type, in particular in work on modeling lexical entry structure to facilitate the development of lexical databases. Bell and Bird (2000) present an early consideration of this topic based on an examination of fifty-five dictionaries and lexicons from a broad set of languages (see also Ide, Kilgarriff, & Romary 2000 for another relevant early work). Something similar is done in Trippel (2006:46–92) where a greater diversity of types of lexical resources is considered (see also Trippel 2009).

Building on previous work (e.g., Gibbon 2002), Trippel (2006:40–45) breaks down the structure of lexicons—and, by extension, lexical data—into three components: microstructure, mesostructure, and macrostructure. The *microstructure* encompasses the information typically associated with the core of a lexical entry (e.g., an orthographic representation of a word, part of speech, and description of meaning). The *macrostructure* corresponds roughly to what, in visual terms, one might refer to as the "layout" of a lexical resource, covering, for instance, how entries are ordered (if ordering is relevant, as is the case for print dictionaries but not necessarily lexical databases), which part of an entry will be privileged for operations such as sorting or referring to an entry (most typically in the form of a headword), and how inflectionally related forms are handled in the lexicon structure. The *mesostructure* is the least prominent aspect of lexicon structure, at least from a presentational standpoint, and it encompasses the various ways that entries can be related to each other (e.g., via cross-references), the nature of the categories used in the lexicon (e.g., transcription conventions or how different subcategories of parts of speech relate to each other), and references to relevant external resources such as a corpus. The fact that lexical data structures show this degree of complexity is a reflection of the fact that analyses of the abstract lexicons associated with languages are themselves complex, requiring reference to a lexicon as a whole, information about individual lexical items, and semantic and formal connections among them.

To make the discussion more concrete, consider the representations of the same lexical information provided in example (4) and figure 3.4, drawn from the TEI Consortium (2019:section 9.3.4).[7] These are partial representations of the information found in a dictionary entry. In (4), a standard presentation format is presented of the sort associated with a print dictionary. In figure 3.4, a partial XML representation, focusing on the etymological content of the entry, is given (see section 3.2 for discussion of XML).[8]

(4) **neume** \'**n(y)üm**\ **n** [F, fr. ML pneuma, neuma, fr. Gk pneuma breath—more at **pneumatic**]: any of various symbols used in the notation of Gregorian chant . . .

```
<entry>
<!—...—>
 <etym>
  <lang>F</lang> fr. <lang>ML</lang>
  <mentioned>pneuma</mentioned>
  <mentioned>neuma</mentioned> fr. <lang>Gk</lang>
  <mentioned>pneuma</mentioned>
  <gloss>breath</gloss>
  <xr type="etym">more at <ptr target="#pneumatic"/>
  </xr>
 </etym>
 <sense>
  <def>any of various symbols used in the notation of
Gregorian chant
<!—...—>
  </def>
 </sense>
</entry>
<!—...—>
<entry xml:id="pneumatic">
 <etym>
<!—...—>
 </etym>
</entry>
```

Figure 3.4
An XML representation of a lexical entry, including mesostructural data.

The information in example (4) and figure 3.4 most directly relates to lexical microstructure because it is primarily encoding a lexical entry. It also saliently encodes mesostructural information in the explicit reference to another lexical entry with the headword *pneumatic*. The presentation format achieves this via the phrasing *more at*. The XML representation does this via a "pointer" tag (abbreviated as ptr in figure 3.4), which references another entry—whose content is mostly unspecified in the XML—with the identifier *pneumatic*. Macrostructural relations are implicit in these representations and connect to the overall conventions used for entry layout and structure. The fact that the entry identifier for one of the entries encoded in figure 3.4 is *pneumatic* reveals an aspect of the macrostructure of this resource, namely that the primary reference point for an entry is some kind of citation form (as is typical for most dictionaries designed for human readability). These remarks cover only a small part of the information encoded in the lexical entry, and some sense of its complexity can be seen by simply comparing the representation in (4) with the one in figure 3.4, which attempts to make explicit much of the information implicitly encoded in (4).

While the example illustrated by (4) and figure 3.4 is drawn from a traditional dictionary entry, the scope of possible kinds of lexical data is quite vast. A widely used lexical data type in historical and comparative linguistics, the word list, represents one possible extreme. The lexical information contained in a word list is minimal in nature, consisting generally merely of a form connected to a semantic label (see Poornima & Good 2010). At the other extreme, resources such as the Oxford English Dictionary (OED Online 2019) can be almost encyclopedic in the information provided in their entries.

Lexical data can also be organized in a variety of ways. The traditional thesaurus, for instance, is oriented around concepts rather than forms. In a similar fashion, the widely used WordNet database (Fellbaum 1998) provides thesaurus-like information, though with more precise semantic specification, in a machine-readable form.

The wide range of ways that lexical data can be organized, along with the fact that being able to compare and combine the information in multiple lexical resources can serve important functions, especially in the domain of translation, has led to significant work on developing generalized models for resources containing lexical data. This can be seen quite clearly, for instance, in the development of lexical markup framework, which can be understood as a metamodel for the creation of lexical resources (see Calzolari, Monachini, & Soria 2013). In the present context, what makes this work of particular interest is that it is based on the analysis of linguistic data as a kind of data in and of itself rather than as the representation of some "deeper" linguistic reality.

As with syntagmatic and paradigmatic analyses, discussed in section 3.3, lexical analyses are, in principle, abstract in nature and can be expressed in various forms, including via annotations, as evidenced by the presentation in figure 3.4.

4 Generalizations about languages and language

The goal of most linguistic scholarship is to discover and make use of generalizations about specific languages or language in general on the basis of data of the kind discussed in sections 2 and 3. To pick some simplistic examples, on the basis of the annotated sentence found in (1), a syntactician might conclude that Yeri is a language where subjects generally precede verbs, or, on the basis of an examination of the oppositions presented in figure 3.1, a phonologist might conclude that Kpelle is a language with a two-way underlying tonemic distinction.

In the present context, these kinds of generalizations are of interest for two reasons. On the one hand, they show how certain kinds of data (e.g., annotations or transcriptions) can be used as the basis for more general claims about a language. Descriptive and formal linguistic analysis, in fact, relies on this kind of analytical step. On the other hand, generalizations such as these can also become data for other kinds of investigation. This is seen especially clearly in the subfield of linguistic typology, which investigates crosslinguistic patterns of variation and, therefore, makes use of language-specific generalizations as a prerequisite to typological investigation.

In figure 3.5, a map presenting the global distribution of basic clausal word order is presented (Dryer 2013). Accompanying this map, table 3.2 indicates how the different symbols on the map translate to specific word order types (where S = subject, O = object, and V = verb) and the number of languages in the sample analyzed as attesting each of the seven types used in the study.

A study like Dryer's can be used to arrive at various kinds of linguistic generalizations, such as observations that SOV and SVO order are by the most common attested

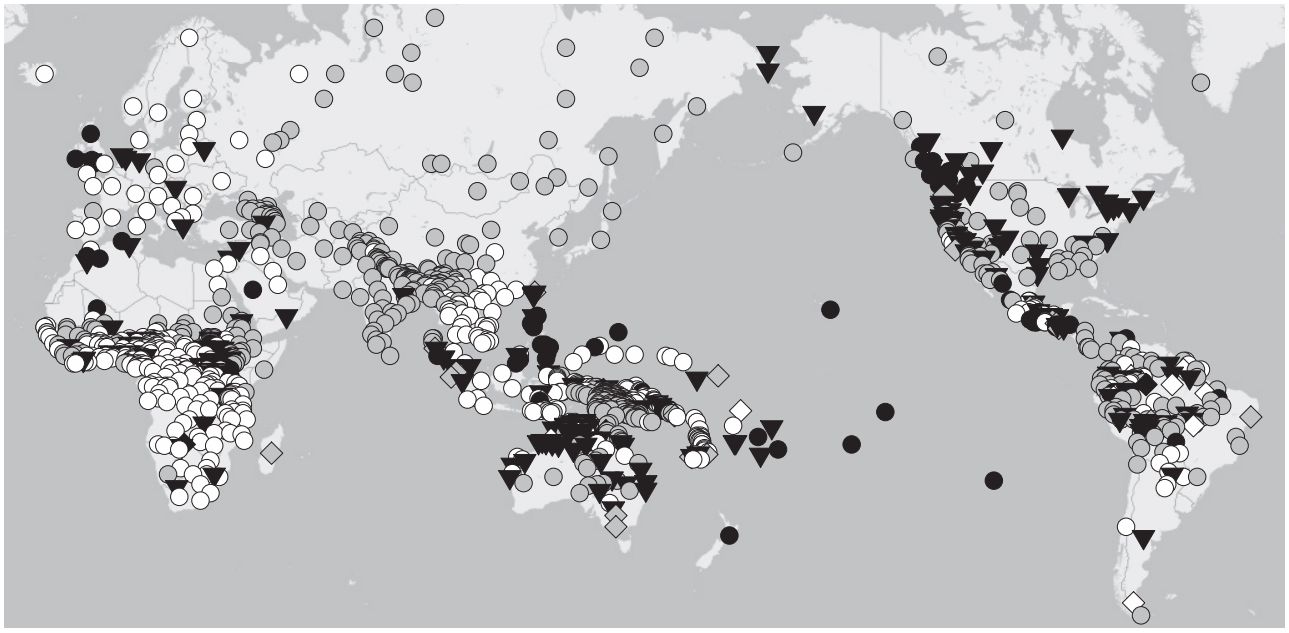


Figure 3.5
The global distribution of different patterns of basic clausal word order.

Table 3.2
Number of languages showing each word order type in a sample of 1,377 languages

Symbol	Basic order	Number
●	SOV	565
○	SVO	488
●	VSO	95
◆	VOS	25
◇	OVS	11
◆	OSV	4
▼	No dominant order	189

basic clausal word order patterns or that SVO word order predominates in sub-Saharan Africa, Southeast Asia, and Europe. These high-level generalizations can only be made when the lower-level generalizations categorizing each language into a specific type are treated as a kind of data. There is thus a kind of analytical "chain" from making a record of observable linguistic behavior, to adding annotations to that record, to devising a general analysis of the characteristics of a language, to assembling those analyses to arrive at generalizations over large classes of languages or, potentially, language in general. Linguistic analysis at one point in the chain can serve as the underlying data for another kind of analysis at a later point in the chain. There is no obvious end point in this kind of data chaining because each new set of generalizations can be the input to further analysis. Higher levels in these chains frequently involve working with quantitative data derived from data collected from observable behavior, and, depending on the point of view of the researcher, the data created at a higher level may be viewed as a "processed" form of the data at a lower level.

There is nothing specifically linguistic about this stepwise pattern of analysis. However, it nevertheless merits consideration in a linguistic context because the kinds of data that will be involved at each step of the chain are specifically linguistic. Fully assessing the range of such chains of data used in the field is outside of the present work, but this would seem to represent an important problem for fully understanding general requirements for linguistic data management. An especially salient division in linguistics in this respect is when and how observable linguistic data come to be used to arrive at generalizations of how language is used in the world, as depicted in figure 3.5, and when and how it is used to motivate proposals for abstract cognitive models of the grammatical knowledge held by language users of the sort associated with the generative tradition.

Perhaps the most important and underappreciated class of linguistic generalizations that are used as data

are the (usually implicit) claims that specific languages exist in the first place. Resources such as Ethnologue (Eberhard, Simons, & Fennig 2019) and Glottolog (Hammarström, Forkel, & Haspelmath 2019) are based on a tremendous amount of data of varying kinds and provide a crucial kind of analytical "infrastructure" for linguistics. Being able to generate a map such as the one in figure 3.5, for instance, presupposes that there is an agreed set of languages in the world and that each can be assigned some location, and typological work generally relies on pre-existing language catalogs and uses their information as data. Resources that aim to be comprehensive on a global scale such as Ethnologue and Glottolog can also be used in studies seeking to understand the state of the world's languages or language families across some significant dimension of variation (e.g., endangerment as in Whalen & Simons 2012). Cysouw and Good (2013) present a proposal on how languages can be defined in terms of the data sets taken to document and describe them, which provides a model for understanding how language catalogs can be created in a way that makes clear the data and analyses that they are based on.

5 Data on language users and situations

Work in sociolinguistics and anthropological linguistics emphasizes the importance of considering data on language use from the perspective of the identity of the participants involved in a given linguistic interaction and the overall context in which the interaction takes place (see, e.g., Hymes [1962] 1971 and Eckert 2012, as well as Poplack, chapter 16, this volume, and Grama, chapter 17, this volume). While detailed sociolinguistic and contextual information is not considered relevant for the interpretation of certain kinds of data (e.g., grammatical acceptability judgments), it is crucial for sociolinguistic investigations and analyses of the relationship between language and culture. Yaeger-Dror and Cieri (2014:467–468) discuss some kinds of identity information that one might need to gather data on to conduct sociolinguistic studies, placing them under the heading of *demographic factors*. These include such things as age, ethnicity, or religion. It may also be important to collect data on the social networks of individuals, and this requires, among other things, collecting identifiers for them (e.g., in the form of a name or a research study code). The precise factors one may want to gather data on are dependent on the study as well as the social characteristics of the individuals and communities being investigated (see, e.g., Stanford & Preston 2009:6–7 for relevant discussion). Work in the variationist sociolinguistic tradition has tended to emphasize the collection of information on demographic factors due to its interest in exploring correlations between aspects of individual identity and language use.

In addition to data about individuals, sociolinguistic and anthropological linguistic research may also collect data on *situational factors* (Yaeger-Dror & Cieri 2014:468) such as the relationship among the people present during a particular speech event, the nature of the place (e.g., public or private) where the event takes place, or how the event fits into social categories of kinds of social situations (e.g., a religious ceremony or casual interaction). Work in the anthropological linguistic tradition often emphasizes detailed consideration of the role of situational factors for influencing language use, and the best-known example of this is probably the line of research falling under the heading of the ethnography of communication (see, e.g., Hymes [1962] 1971 for a foundational work and Michael 2011:126–128 for overview discussion). To give some sense of the range of situational factors that might need to be considered as well as their potential cultural specificity, the study by Duranti (1981:361–363) of language use in village council meetings within a Samoan village identified the following as significant for understanding the characteristics of turn-taking during these events: the seating arrangement of the participants, the order in which participants are served a ceremonially important drink, the position of a participant's legs while sitting, and whether a participant is wearing clothing on the upper part of their body.

While data about language users and the situations in which speech events take place is not linguistic data in a narrow sense in that they are not strictly about language itself, they are clearly linguistic data in the broader sense that they are collected and used by linguists to come to a better understanding of data derived from observable linguistic behavior.

6 Linguistic metadata, data preservation, and data replicability

While not central to traditional linguistic research, the rise of digital language resources has led to metadata becoming an increasingly important kind of data for linguistics

(as well as many other fields). *Metadata* is generally defined along the lines of data about data, though, in the present context, some qualification of this definition is in order. There are some kinds of data about data that are best considered to simply be new data formed on the basis of existing data. This is the case, for instance, with many of the kinds of annotations discussed in section 3.2. The term metadata is generally applied to data about other data when the new data are not seen as directly supporting further analysis. In example (1), the identifier included with the example, 120517–001, which is used to associate the transcription with the recording that it is based on, would normally be considered metadata because an identifier that serves a largely "bookkeeping" function is not treated as new data that support further linguistic analysis.

In some cases, whether a piece of information would be considered metadata or not is dependent on the kind of analysis being conducted. Again, looking at (1), it includes an identifier of the speaker, JS, who produced the utterance. If this example were being used to illustrate a typological feature of the language, then this identifier would be most readily classified as metadata. If this example were being used in a sociolinguistic study to illustrate variation of some kind among different language users, then this identifier would move closer to being data. Strictly speaking, the identifier itself would still be metadata because it is merely an identifier for a person, not the person itself. However, a key difference is that the identity of the speaker could be relevant for understanding variation in a language but would not be relevant if the example is being used to illustrate something understood as a general fact of the language.

If we understand *linguistic* data to be data about languages, then metadata would not, strictly speaking, be linguistic data. If we think of linguistic data as the data needed to effectively conduct linguistic research, then metadata are taking on an increasingly important role in any kind of research that involves the collection of resources that are considered valuable for long-term preservation and of interest beyond the specific project for which they were collected (see also Andreassen, chapter 7, this volume). Resources produced during the course of endangered language documentation (see section 2.2) fall into this class as do resources associated with a project like TalkBank (MacWhinney 2007), which seeks to support data sharing among researchers collecting recordings of conversational interactions.

Broadly speaking, an important division here is between data produced under a paradigm of replicability as opposed to reproducibility (see Berez-Kroeker et al. 2017 for detailed consideration in a linguistic context as well as Gawne & Styles, chapter 2, this volume). Replicable research methods, in principle, allow for equivalent data to be collected at more than one place and time so that results can be verified across multiple studies. Reproducible research methods allow for access to the data on which a study was originally based and provide enough details on the methodology used to analyze that data so that another researcher can verify how the results of the original research were arrived at (Berez-Kroeker et al. 2017:4–5). Linguistic research will never permit the same degree of replicability as is possible in some of the natural sciences. However, some methodological approaches, such as the collection of grammaticality judgments for a given language based on a carefully prepared list of sentences or the phonetic analysis of words collected in a highly controlled way, should allow for replication assuming similar kinds of language users can be found. Other approaches, such as the collection of an oral history from a skilled storyteller of an endangered language or child language data recorded within a household, do not allow for replicability. This division is not necessarily a strict one. Data gathered via the use of a standard prompt across subjects, such as the well-known example of the so-called frog stories where language users are shown a picture-based narrative and asked to recount it orally (see Berman & Slobin 1994), can be expected to be broadly similar across studies if collected across a wide enough pool of language users, but significant individual-level variation would also be considered normal in a way that would be less expected in the case of, for instance, a carefully controlled grammatical judgment elicitation task.

While modern data storage technologies greatly facilitate the preservation of all kinds of linguistic data, long-term archiving has been the highest priority for data that is either difficult or impossible to replicate, whether because there are significant barriers to collecting it more than once (e.g., a recording from one of the last users of an endangered language) or because it requires significant resources to create (e.g., large annotated text corpora). Data of these kinds have been the primary focus for the development of metadata standards for linguistics because of the likelihood that they

will be reused. The two sets of metadata standards that are probably most widely discussed within linguistics at present are the Open Language Archives Community (OLAC) standard (Simons & Bird 2008) and the Component Metadata Infrastructure (CMDI) framework (Broeder et al. 2012). The OLAC standard, by design, provides a specification of the baseline metadata that should be associated with a linguistic resource, and this makes it optimized for information exchange about the content of an archive, for example, to support resource discovery. CMDI has a component-based model that allows different projects to create their own metadata standards by using parts of other standards or defining specific sets of metadata fields for their own needs.

Metadata standards developed within the CMDI framework are designed to be as compatible with each other as their different content will allow. The framework was created in recognition of the fact that the specific metadata needs of different subfields are too diverse to be subsumed under a single standard, and its design, in part, represents a response to issues that arose in developing the ISLE Metadata Initiative (IMDI) standard (Broeder et al. 2012:1), which was primarily intended for use with multimedia corpora (Broeder & Van Uytvanck 2014:159).[9] IMDI is much more expansive than OLAC and a comparison of the two gives some sense for the possible scope of linguistic metadata (see also Austin 2006:94 for a list of possible functions for linguistic metadata). OLAC metadata are primarily oriented around creating a basic descriptive record for each linguistic resource in a collection, while IMDI metadata additionally allow for the description of information about the relationships among resources within a collection to be described (e.g., grouping a given set of resources into a corpus corresponding to work done during a specific project), the characteristics of the people involved in the creation of a resource, and the context in which a recording was made, among other kinds of information (see Broeder & Wittenburg 2006:127).

It seems clear that linguistic metadata will increase in importance as a kind of linguistic data as more resources become digitally available and expectations increase for their degree of interoperation, that is, the ability for the information contained within them to be effectively combined for different applications (see Witt et al. 2009 for relevant discussion in a linguistic context and Wilkinson et al. 2016 for more general discussion). Finally, while the discussion in this chapter is largely aimed at the research community, it is worth emphasizing that, in many cases, languages resources will be of interest to language communities as well as the general public, and metadata can play a central role in ensuring that those resources can be discovered and used in a wide range of contexts.

7 The breadth and expanding scope of linguistic data

While "data" play a leading role in analysis across all subdisciplines of linguistics, understanding the full scope of the data made use of by linguists has not been a central concern of the field. Nevertheless, the wide range of data types involved in linguistic analysis would seem to provide an underrecognized opportunity for linguistics to be at the forefront of questions of data management because linguists are familiar with a much wider range of data types than are scholars working in many other disciplines.

In fact, if one understands linguistic data to include data of any kind that are used to support linguistic investigation, this survey has, for reasons of practicality, left out many kinds of data that are not specifically connected to language but are important in interdisciplinary studies. Bostoen et al. (2015), for example, place linguistic data alongside data from biogeography, palynology (the study of particulate samples), and archaeology in making a proposal regarding the dynamics of the Bantu language expansion in sub-Saharan Africa. Similarly, Pakendorf et al. (2017) consider the relationship between linguistic and genetic data as a means of arriving at a better understanding of the historical forces shaping language contact among different language groups in southern Africa. In other domains, Yu, Abrego-Collier, and Sonderegger (2013) combine linguistic data with information from psychological assessment tasks to look at factors that could explain individual differences in language use, and Berez (2015) demonstrates how the integration of linguistic data and geographic information systems data can yield useful insights in the analysis of spatial language. If we include data from allied fields alongside "core" kinds of data about language, then it is clear that the scope of "linguistic" data can extend far beyond what has been discussed here.

As a final note, it seems worth briefly remarking on the interplay between data and the methods used to analyze data. Heggarty, Maguire, and McMahon (2010), for instance, consider one of the most long-standing

problems in linguistic analysis—the reconstruction of the historical relationships of languages within a language family—in light of the availability of phylogenetic methods and tools developed within the biological sciences. The use of these methods raises significant questions regarding the curation and coding of lexical data (e.g., whether binary or multistate variables should be used) that were not in focus when more traditional methods were employed.

Time-aligned annotation of the sort discussed in section 3.2 provides another relevant example. The availability of technologies allowing for the creation of digital recordings and tools to annotate those recordings has, in effect, created a new kind of linguistic data that did not exist previously. The field of linguistics is currently seeing rapid changes in the methods used to analyze language, and it seems likely that in the coming years this will result in the scope of linguistic data being similarly expanded to better support the kinds of analyses these methods support.

Notes

1. I would like to thank the editors of this Handbook as well as Philipp Conzett and Koenraad De Smedt for comments on an earlier version of this chapter.

2. "Hybrid" corpora also exist that combine primary text data with data in other formats such as audio and video with accompanying transcription (see, e.g., Anderson, Beavan, & Kay 2007).

3. I use the term acceptable here broadly to encompass sentences that are considered syntactically grammatical, semantically interpretable (rather than anomalous), or pragmatically felicitous.

4. See section 4 for discussion of Glottolog (Hammarström, Forkel, & Haspelmath 2019), which is the source of glottocodes.

5. The abbreviations in the gloss line of the example in (1) are used as follows: 1 = first person, 2 = second person, F = feminine, R = realis, SG = singular.

6. The full metadata record for the recording associated with the example in (1) can be found at https://hdl.handle.net/1839/B7031E9D-78CC-44DF-8B2A-889D541E3180.

7. The example in TEI Consortium (2019:section 9.3.4) appears to be adapted from an edition of Merriam-Webster's *New Collegiate Dictionary*. The example uses a markup format developed by the Text Encoding Initiative, which is more widely used in fields outside of linguistics. However, see Romary and Witt (2014) for consideration of the application of the recommendations of this initiative to linguistic resources.

8. The sequence <!— . . . —> in the XML representation in figure 3.4 is a so-called comment, meaning that it is to be interpreted as a human-readable remark on the XML structure. In this case, the comment consists of ". . ." and indicates that some aspects of the data represented in (4) are left out of the XML representation.

9. Detailed discussion of the IMDI standard is found in Broeder and Wittenburg (2006).

References

Abeillé, Anne. 2003. Introduction. In *Treebanks: Building and Using Parsed Corpora*, ed. Anne Abeillé, xiii–xxvi. Dordrecht: Kluwer.

Abrusán, Márta. 2019. Semantic anomaly, pragmatic infelicity, and ungrammaticality. *Annual Review of Linguistics* 5 (1): 329–351.

Anderson, Jean, Dave Beavan, and Christian Kay. 2007. SCOTS: Scottish Corpus of Texts and Speech. In *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, ed. Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, 17–34. Basingstoke, UK: Palgrave Macmillan.

Austin, Peter K. 2006. Data and language documentation. In *Essentials of Language Documentation*, ed. Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, 87–112. Berlin: Mouton de Gruyter.

Baerman, Matthew, Dunstan Brown, and Greville G. Corbett. 2005. *The Syntax–Morphology Interface: A Study of Syncretism*. Cambridge: Cambridge University Press.

Beal, Joan C., Karen P. Corrigan, and Hermann L. Moisl, eds. 2007. *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*. Basingstoke, UK: Palgrave Macmillan.

Bell, John, and Steven Bird. 2000. A preliminary study of the structure of lexicon entries. In *Proceedings from the Workshop on Web-Based Language Documentation and Description*. Philadelphia, December 12–15, 2000. https://web.archive.org/web/20010603214720/https://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html (originally available at http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html).

Bender, Emily M., Sumukh Ghodke, Timothy Baldwin, and Rebecca Dridan. 2012. From database to treebank: On enhancing hypertext grammars with grammar engineering and treebank search. *Language Documentation and Conservation* Special Publication 4:179–206. http://hdl.handle.net/10125/4535.

Beres, Anna M. 2017. Time is of the essence: A review of electroencephalography (EEG) and event-related brain potentials (ERPs) in language research. *Applied Psychophysiology and Biofeedback* 42 (4): 246–255. https://doi.org/10.1007/s10484-017-9371-3.

Berez, Andrea L. 2015. Directionals, episodic structure, and geographic information systems: AREA/PUNCTUAL distinctions in Ahtna travel narration. *Linguistics Vanguard* 1 (1): 155–175.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2017. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18.

Berman, Ruth A., and Dan Isaac Slobin. 1994. *Relating Events in Narrative: A Cross-linguistic Developmental Study*. Hillsdale, NJ: Lawrence Erlbaum.

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath. 2008. *The Leipzig Glossing Rules: Conventions for Interlinear Morpheme by Morpheme Glosses*. https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf.

Birch, Bruce. 2014. Data collection. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 27–45. Oxford: Oxford University Press.

Bird, Steven. 1999. Multidimensional exploration of online linguistic field data. In *Proceedings of the 29th Annual Meeting of the Northeast Linguistics Society*, ed. Pius Tamanji, Masako Hirotani, and Nancy Hall, 33–47. Amherst, MA: Graduate Linguistics Student Association.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79:557–582.

BNC Consortium. 2007. *The British National Corpus,* version 3 (BNC XML edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Boas, Franz. 1911. Linguistics and ethnology. In *Handbook of American Indian languages*, ed. Franz Boas, 59–73. Washington, DC: Government Printing Office.

Boersma, Paul, and David Weenink. 2019. *Praat: Doing Phonetics by Computer* [computer program]. http://www.praat.org/.

Bonelli, Elena Tognini. 2010. Theoretical overview of the evolution of corpus linguistics. In *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and Michael McCarthy, 14–27. Abingdon, UK: Routledge.

Bostoen, Koen, Bernard Clist, Charles Doumenge, Rebecca Grollemund, Jean-Marie Hombert, Joseph Koni Muluwa, and Jean Maley. 2015. Middle to late Holocene paleoclimatic change and the early Bantu expansion in the rain forests of Western Central Africa. *Current Anthropology* 56 (3): 354–384.

Bow, Catherine, Baden Hughes, and Steven Bird. 2003. Towards a general model for interlinear text. In *Proceedings of E-MELD 2003: Digitizing and Annotating Texts and Field Recordings*. East Lansing, Michigan, July 11–13. http://e-meld.org/workshop/2003/bowbadenbird-paper.html.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Oxford: Blackwell.

Broeder, Daan, and Dieter Van Uytvanck. 2014. Metadata formats. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 150–165. Oxford: Oxford University Press.

Broeder, Daan, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: A component metadata infrastructure. In *Proceedings of the Workshop on Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR, LREC 2012,* Istanbul, Turkey, 1–4. http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf.

Broeder, Daan, and Peter Wittenburg. 2006. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies* 1 (2): 119–132.

Brugman, Hennie, and Albert Russel. 2004. Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2065–2068. Lisbon: ELRA. http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf.

Calzolari, Nicoletta, Monica Monachini, and Claudia Soria. 2013. LMF: Historical context and perspectives. In *LMF: Lexical Markup Framework*, ed. Gil Francopoulo, 1–18. London: ISTE Ltd and Wiley and Sons.

Carpenter, Bob. 1992. *The Logic of Typed Feature Structures*. Cambridge: Cambridge University Press.

Childs, G. Tucker, Jeff Good, and Alice Mitchell. 2014. Beyond the ancestral code: Towards a model for sociolinguistic language documentation. *Language Documentation and Conservation* 8: 168–191.

Cohn, Abigail. 2001. Phonology. In *The Handbook of Linguistics*, ed. Mark Aronoff and Janie Rees-Miller, 180–212. Oxford: Blackwell.

Cysouw, Michael, and Jeff Good. 2013. Languoid, doculect, and glossonym: Formalizing the notion "language." *Language Documentation and Conservation* 7:331–359.

Davies, Mark. 2008–. *The Corpus of Contemporary American English (COCA): 600 million words, 1990–present*. https://corpus.byu.edu/coca/.

Davies, Mark. 2013. *Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day*. https://corpus.byu.edu/now/.

Diessel, Holger. 2017. Usage-based linguistics. *Oxford Research Encyclopedia of Linguistics*. https://dx.doi.org/10.1093/acrefore/9780199384655.013.363.

Diestel, Reinhard. 1997. *Graph Theory*. New York: Springer.

Dimmendaal, Gerrit J. 2010. Language description and "the new paradigm": What linguists may learn from ethnocinematographers. *Language Documentation and Conservation* 4:152–158.

Dryer, Matthew S. 2013. Order of subject, object and verb. In *The World Atlas of Language Structures Online*, ed. Matthew S.

Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/81.

Duranti, Alessandro. 1981. Speechmaking and the organisation of discourse in a Samoan *fono*. *Journal of the Polynesian Society* 90 (3): 357–400.

Durkin, Philip, ed. 2016. *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, eds. 2019. *Ethnologue: Languages of the World*, 22nd ed. Dallas: SIL International. http://www.ethnologue.com.

Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41 (1): 87–100.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Gibbon, Dafydd. 2002. Prosodic information in an integrated lexicon. In *Speech Prosody 2002*. ISCA Archive. http://www.isca-speech.org/archive/sp2002.

Gippert, Jost. 2006. Linguistic documentation and the encoding of textual materials. In *Essentials of Language Documentation*, ed. Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, 337–361. Berlin: Mouton de Gruyter.

Good, Jeff. 2004. The descriptive grammar as a (meta)database. In *Proceedings of E-MELD 2004: Linguistic Databases and Best Practice*. Detroit, Michigan. July 15–18. http://www.e-meld.org/workshop/2004/jcgood-paper.html.

Good, Jeff. 2011. Data and language documentation. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 212–234. Cambridge: Cambridge University Press.

Goodman, Michael Wayne, Joshua Crowgey, Fei Xia, and Emily M. Bender. 2015. Xigt: Extensible interlinear glossed text for natural language processing. *Language Resources and Evaluation* 49 (2): 455–485.

Gries, Stefan Th., and Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In *Handbook of Linguistic Annotation*, ed. Nancy Ide and James Pustejovksy, 379–409. Dordrecht, the Netherlands: Springer.

Grieve, Jack, Andrea Nini, and Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* 21 (1): 99–127.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2019. *Glottolog 3.4*. Jena, Germany: Max Planck Institute for the Science of Human History. https://glottolog.org/.

Hayes, Bruce. 1990. Precompiled phrasal phonology. In *The Phonology-Syntax Connection*, ed. Sharon Inkelas and Draga Zec, 85–108. Stanford: CSLI.

Heggarty, Paul, Warren Maguire, and April McMahon. 2010. Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:3829–3843.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161–195.

Himmelmann, Nikolaus P. 2012. Linguistic data types and the interface between language documentation and description. *Language Documentation and Conservation* 6:187–207.

Hyman, Larry M. 2011. Tone: Is it different? In *The Handbook of Phonological Theory*, 2nd ed., ed. John A. Goldsmith, Jason Riggle, and Alan C. L. Yu, 197–239. Chichester: Wiley-Blackwell.

Hymes, Dell H. (1962) 1971. The ethnography of speaking. In *Anthropology and Human Behavior*, ed. Thomas Gladwin and William C. Sturtevant, 13–53. Washington, DC: The Anthropological Society of Washington.

Ide, Nancy. 2017. Introduction: The handbook of linguistic annotation. In *Handbook of Linguistic Annotation*, ed. Nancy Ide and James Pustejovksy, 1–18. Dordrecht, the Netherlands: Springer.

Ide, Nancy, Christian Chiarcos, Manfred Stede, and Steve Cassidy. 2017. Designing annotation schemes: From model to representation. In *Handbook of Linguistic Annotation*, ed. Nancy Ide and James Pustejovksy, 73–111. Dordrecht, the Netherlands: Springer.

Ide, Nancy, Adam Kilgarriff, and Laurent Romary. 2000. A formal model of dictionary structure and content. In *Proceedings of the Ninth EURALEX International Congress*, ed. Ulrich Heid, Stefan Evert, Egbert Lehmann, and Christian Rohrer, 113–126. Stuttgart, Germany: Institut für Maschinelle Sprachverarbeitung. http://arxiv.org/abs/0707.3270.

Ide, Nancy, and James Pustejovksy, eds. 2017. *Handbook of Linguistic Annotation*. Dordrecht, the Netherlands: Springer.

Kaan, Edith. 2007. Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass* 1 (6): 571–591.

Kendall, Tyler. 2008. On the history and future of sociolinguistic data. *Language and Linguistics Compass* 2 (2): 332–351.

Kendall, Tyler. 2011. Corpora from a sociolinguistic perspective. *Revista Brasileira de Linguística Aplicada* 11:361–389. doi:10.1590/S1984–63982011000200005.

Koenig, W., H. K. Dunn, and L. Y. Lacy. 1946. The sound spectrograph. *Journal of the Acoustical Society of America* 18:19–49. doi:10.1121/1.1916342.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Ladefoged, Peter. 1957. Use of palatography. *Journal of Speech and Hearing Disorders* 22 (5): 764–774.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites*. Stanford: Stanford University Press.

MacWhinney, Brian. 2007. The TalkBank project. In *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, ed. Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, 164–180. Basingstoke, UK: Palgrave Macmillan.

Margetts, Anna, and Andrew Margetts. 2012. Audio and video recording techniques for linguistic research. In *The Oxford Handbook of Linguistic Fieldwork*, ed. Nicholas Thieberger, 13–53. Oxford: Oxford University Press.

Maxwell, Mike. 2012. Electronic grammars and reproducible research. *Language Documentation and Conservation* Special Publication 4:207–234. http://hdl.handle.net/10125/4536.

McCarthy, Michael, and Anne O'Keeffe. 2010. Historical perspective: What are corpora and how have they evolved? In *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and Michael McCarthy, 3–13. Abingdon, UK: Routledge.

McCawley, James D. 1982. Parentheticals and discontinuous constituent structure. *Linguistic Inquiry* 13:91–106.

McCawley, James D. 1998. *The Syntactic Phenomena of English*, 2nd ed. Chicago: University of Chicago Press.

Michael, Lev. 2011. Language and culture. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 120–140. Cambridge: Cambridge University Press.

Moravcsik, Edith A. 2010. Conflict resolution in syntactic theory. *Studies in Language* 34 (3): 636–669.

Mosel, Ulrike. 2006. Grammaticography: The art and craft of writing grammars. In *Catching Language: The Standing Challenge of Grammar Writing*, ed. Felix Ameka, Alan Dench, and Nicholas Evans, 41–68. Berlin: Mouton de Gruyter.

Nadeau, David, and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes* 30 (1): 3–26.

Newmeyer, Frederick J. 1998. *Language Form and Language Function*. Cambridge: MIT Press.

Nordhoff, Sebastian. 2012. The grammatical description as a collection of form-meaning-pairs. *Language Documentation and Conservation* Special Publication 4:33–62. http://hdl.handle.net/10125/4529.

OED Online. 2019. Oxford: Oxford University Press. http://www.oed.com.

Pakendorf, Brigitte, Hilde Gunnink, Bonny Sands, and Koen Bostoen. 2017. Prehistoric Bantu-Khoisan language contact. *Language Dynamics and Change* 7 (1): 1–46.

Palmer, Alexis, and Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed text. In *Proceedings of the Linguistic Annotation Workshop (LAW '07)*, 176–183. Stroudsburg, PA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1642059.1642087.

Penton, David, Catherine Bow, Steven Bird, and Baden Hughes. 2004. Towards a general model for linguistic paradigms. In *Proceedings of E-MELD 2004: Linguistic Databases and Best Practice*. Detroit, Michigan, July 15–18. http://e-meld.org/workshop/2004/bird-paper.pdf.

Phillips, Colin, and Matthew Wagers. 2007. Relating structure and time in linguistics and psycholinguistics. In *The Oxford Handbook of Psycholinguistics*, ed. M. Gareth Gaskell, 739–756. Oxford: Oxford University Press.

Poornima, Shakthi, and Jeff Good. 2010. Modeling and encoding traditional wordlists for machine applications. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground (NLPLING '10)*, ed. Fei Xia, William Lewis, and Lori Levin, 1–9. Stroudsburg, PA: Association for Computational Linguistics.

Pustejovksy, James, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. In *Handbook of Linguistic Annotation*, ed. Nancy Ide and James Pustejovksy, 21–72. Dordrecht, the Netherlands: Springer.

Romary, Laurent, and Andreas Witt. 2014. Data formats for phonological corpora. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 166–190. Oxford: Oxford University Press.

Sag, Ivan A., Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*, 2nd ed. Stanford: CSLI.

Schembri, Adam. 2010. Documenting sign languages. In *Language Documentation and Description,* vol. 7, ed. Peter K. Austin, 105–143. London: SOAS.

Schultze-Berndt, Eva. 2006. Linguistic annotation. In *Essentials of Language Documentation*, ed. Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, 213–251. Berlin: Mouton de Gruyter.

Schütze, Carson T. (1996) 2016. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Berlin: Language Science Press.

Sebba, Mark, and Susan Dray. 2007. Developing and using a corpus of written creole. In *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, ed. Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, 181–204. Basingstoke, UK: Palgrave Macmillan.

Seyfeddinipur, Mandana. 2012. Reasons for documenting gesture and suggestions for how to go about it. In *The Oxford Handbook of Linguistic Fieldwork*, ed. Nicholas Thieberger, 147–165. Oxford: Oxford University Press.

Seyfeddinipur, Mandana, and Felix Rau. Keeping it real: Video data in language documentation and language archiving. *Language Documentation and Conservation* 14:503–519.

Simons, Gary F. 2006. *Ensuring That Digital Data Last: The Priority of Archival Form over Working Form and Presentation Form*. Dallas: SIL International. https://web.archive.org/web/20130315021812/http://www-01.sil.org/silewp/2006/003/SILEWP2006-003.htm (originally available at: http://www.sil.org/silewp/2006/003/SILEWP2006-003.htm).

Simons, Gary F., and Steven Bird, eds. 2008. *OLAC Metadata*. http://www.language-archives.org/OLAC/metadata.html.

Sprouse, Jon. 2013. Acceptability judgments. In *Oxford Bibliographies Online: Linguistics*. Oxford: Oxford University Press.

Stanford, James N., and Dennis R. Preston. 2009. The lure of a distant horizon: Variation in Indigenous minority languages. In *Variation in Indigenous Minority Languages*, ed. James N. Stanford and Dennis R. Preston, 1–20. Amsterdam: Benjamins.

Strassel, Stephanie, and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3273–3280. Paris: ELRA.

TEI Consortium. 2019. Dictionaries. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange [version 3.5.0]*, ed. TEI Consortium, chapter 9. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, ed. Patience Epps and Alexandre Arkhipov, 389–407. Berlin: Mouton de Gruyter.

Thieberger, Nicholas, and Andrea L. Berez. 2012. Linguistic data management. In *The Oxford Handbook of Linguistic Fieldwork*, ed. Nicholas Thieberger, 90–118. Oxford: Oxford University Press.

Trippel, Thorsten. 2006. *The Lexicon Graph Model: A Generic Model for Multimodal Lexicon Development*. Saarbrücken, Germany: AQ-Verlag.

Trippel, Thorsten. 2009. Representation formats and models for lexicons. In *Representation Formats and Models for Lexicons*, ed. Andreas Witt and Dieter Metzing, 165–184. Berlin: Springer.

van Marle, Jaap. 2000. Paradigmatic and syntagmatic relations. In *Morphology: An International Handbook on Inflection and Word-formation*, ed. Geert Booij, Christian Lehmann, and Joachim Mugdan, 225–234. Berlin: De Gruyter.

Welmers, William E. 1962. The phonology of Kpelle. *Journal of African Languages* 1:69–93.

Whalen, D. H., and Gary F. Simons. 2012. Endangered language families. *Language* 88 (1): 155–173.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Willems, Roel M., and Marcel A. J. van Gerven. 2018. New fMRI methods for the study of language. In *The Oxford Handbook of Psycholinguistics*, ed. Shirley-Ann Rueschemeyer and M. Gareth Gaskell, 975–991. Oxford: Oxford University Press.

Wilson, Jennifer. 2014. Yeri. *The Language Archive*. https://hdl.handle.net/1839/00-0000-0000-001A-E16A-5.

Wilson, Jennifer. 2017. A grammar of Yeri: A Torricelli language of Papua New Guinea. PhD dissertation, University at Buffalo.

Witt, Andreas, Ulrich Heid, Felix Sasaki, and Gilles Sérasset. 2009. Multilingual language resources and interoperability. *Language Resources and Evaluation* 43 (1): 1–14.

Woodbury, Anthony C. 2011. Language documentation. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 159–186. Cambridge: Cambridge University Press.

Yaeger-Dror, Malcah, and Christopher Cieri. 2014. Introduction to the special issue on archiving sociolinguistic data. *Language and Linguistics Compass* 8 (11): 465–471.

Yu, Alan C. L., Carissa Abrego-Collier, and Morgan Sonderegger. 2013. Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PLOS ONE* 8:e74746.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363