Introduction: The Handbook of Linguistic Annotation

Nancy Ide

1 Introduction

Linguistic annotation of language data was originally performed in order to provide information for the development and testing of linguistic theories, or, as it is known today, corpus linguistics. At the time, considerable time and effort was required to annotate data with even the simplest linguistic phenomena, and the annotated corpora available for study were quite small. Over the past three decades, advances in computing power and storage together with development of robust methods for automatic annotation have made linguistically-annotated data increasingly available in ever-growing quantities. As a result, these resources now serve not only linguistic studies, but also the field of natural language processing (NLP), which relies on linguistically-annotated text and speech corpora to evaluate new human language technologies and, crucially, to develop reliable statistical models for training these technologies. In recent years, there has been a noticeable upswing in linguistic annotation activity, which has expanded to cover a wide variety of linguistic phenomena. The rise in annotation activity has also come with a proliferation of annotation tools to support the creation and storage of labeled data, means for collaborative and distributed annotation efforts, and the introduction of crowdsourcing mechanisms such as Amazon Mechanical Turk.

The goal of this volume is to provide a comprehensive survey of the development and state-of-the-art for linguistic annotation of language resources, including methods for annotation scheme design, annotation creation, physical format considerations, annotation tools, annotation use, evaluation, etc. The volume is divided into two parts: Part I includes survey chapters on the various phases and considerations for an annotation project, and Part II consists of thirty-nine case studies describing major annotation projects for a broad range of linguistic phenomena. The motiva-

Nancy Ide
Department of Computer Science, Vassar College, Poughkeepsie, NY 12604 USA e-mail: ide@cs.vassar.edu

tion for including detailed descriptions of an extensive set of annotation projects is, first, that given the common notion of what comprises a valid or valuable academic contribution, such descriptions are rarely published and therefore very often unavailable. Second, by providing precise descriptions of methods, lessons learned and experience gained, these case studies are likely the most valuable pieces of information to guide those who intend to undertake an annotation project. Thus Parts I and II are intended to be complementary, providing, on the one hand, an overview of what is currently understood to be best practice in the field, and, on the other, a detailed accounting of actual practice over the past several years.

2 A Brief Anatomy of Linguistic Annotation Projects

Linguistic annotation involves the association of descriptive or analytic notations with language data. The raw data may be textual, drawn from any source or genre, or it may be in the form of time functions (audio, video and/or physiological recordings). The annotations themselves may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tags, syntactic analyses, "named entity" labels, semantic role labels, time and event identification, coreference chains, discourse-level analyses, and many others. Resources vary in the range of annotation types they contain: some resources contain only one or two types, while others contain multiple annotation "layers" or "tiers" of linguistic descriptions.

The most critical component of a linguistic annotation project is the *annotation scheme* that defines the labels and associated features to be associated with the appropriate *annotation unit* (e.g., a type of sound, token or word, phrase, clause, document). The labels and units must have operational definitions so that humans looking at the same piece of data are more likely to assign it the same label. Schemes that exist for the purpose of training automatic machine annotators may identify features (e.g., orthographic attributes, ngrams, or information from other annotations such as part of speech, subject/object, semantic role, etc.) that are highly correlated with the annotation labels.

An annotation project may use an existing scheme or it may demand development of a new scheme for phenomena that have not been previously considered. If the latter, the project may spend more time on scheme development than on annotation, whether it is designed *a priori* or developed iteratively with cycles of annotation, evaluation, and revision of the scheme. Finding a balance between a sufficiently rich description of the linguistic phenomenon in question and the ability of humans and/or machines to reliably and consistently identify it is arguably the most important part of an annotation project.

Finally, modern manual or semi-automatic annotation efforts typically rely on an *annotation tool* with an interface that enables identification of spans of characters and/or links between such spans, together with means to associate a label or labels with the identified spans and/or links; this may be accompanied by tools to measure

inter-annotator agreement (IAA) for two or more annotators using one of several popular metrics, in order to measure consensus, define a threshold of expected performance by automatic annotation tools, and/or determine if a particular scale is appropriate for measuring the phenomenon in question, etc.

All of these fundamental components of a linguistic annotation project have undergone significant evolution over the past five decades. The following section outlines the history and evolution of linguistic annotation starting in the mid-twentieth century, and gives pointers to chapters in Part I of this volume that fill in the picture by describing state-of-the-art methods and best practices for linguistic annotation as it is practiced today.

3 History, Evolution, and State-of-the-Art

The first modern, electronically-readable annotated corpus was the one million-word Brown Corpus of Standard American English, which in its original unannotated form served as the basis for Henry Kuĉera and W. Nelson Francis' *Computational Analysis of Present-Day American English* [32]. Over the following decade, in what is arguably the first modern linguistic annotation project, part-of-speech annotation was added to the Brown Corpus, fostering the development of increasingly accurate automatic methods for part-of-speech tagging[1] in order to avoid the painstaking work of manual validation. Like the Brown Corpus, corpora developed in the 70s and 80s were typically annotated for part-of-speech, but the lack of reasonably accurate automatic methods and the high cost of manual annotation disallowed the production of sufficiently large corpora containing annotations for other linguistic phenomena, such as syntax.

In the late 1980s, the new availability large-scale language data led to a proliferation of linguistic annotation projects, most focused on part-of-speech (or richer morpho-syntactic) annotations, and spearheaded the use of probabilistic methods for automatic annotation based on statistical data derived from the corpus. The first major effort of this kind produced morpho-syntactic and syntactic annotations of the one-million-word Lancaster-Oslo-Bergen (LOB) corpus of English [18]. Building on this work, the Penn Treebank project [37] produced a one-million-word corpus of *Wall Street Journal* articles annotated for part-of-speech and skeletal syntactic annotations and, later, basic functional information [36]. Automatically-produced annotations subsequently validated by humans (in whole or in part) were used to create several other major corpora in the 1990s, including the 100-million word British National Corpus [7], released in 1994; corpora produced by the MULTEXT project (1993-96) [29] and its follow-on, MULTEXT-EAST (1994-97) [15], which provided parallel aligned corpora in a dozen Western and Eastern languages anno-

[1] The earliest automatic part-of-speech taggers include Greene and Rubin's TAGGIT [19], Garside's CLAWS [17], DeRose's VOLSUNGA [13], and Church's PARTS [6]

tated for part-of-speech; and the PAROLE and SIMPLE corpora[2], which included part-of-speech tagged data in fourteen European languages.

Speaking broadly, annotation projects undertaken in the 1990s share some common characteristics. One is methodology: by far the most common strategy was the automatic generation of annotations that were subsequently validated by humans.[3] Since 2000, annotation methodology has expanded to include strategies such as pair annotation (see Demirşahin and Zeyrek, Part II, IV.b.ii) and iterative enhancement (Dickinson and Tufiş, Part I.V.e) based on error detection. The most notable development is the attempt to defray the high cost of annotated resource development through *crowdsourcing* using Amazon Mechanical Turk[4] and similar systems, and the so-called "games-with-a-purpose", as described in Poesio *et al.* (Chapter V.f).

Another commonality among projects in the 1990s is, in fact, the lack of commonality among these projects, in terms of both the physical formats used to represent the annotated data[5] and the linguistic labels used in the annotation schemes. In Europe, the need to harmonize annotated resources across multiple languages led to the development of standards for linguistic annotations in the EU-funded EAGLES project[6], whose guidelines were followed in major EU resource development projects such as MULTEXT, MULTEXT-EAST, and PAROLE/SIMPLE. EAGLES published an influential set of tiered specifications for morpho-syntactic annotation for multiple languages[7] and the encoding of document structure and basic linguistic elements in linguistically-annotated corpora [24, 27]. Standards for annotating speech phenomena such as prosody were also proposed at this time (e.g., [47]). Apart from these efforts, which were known and used primarily in Europe, few guidelines or standards for linguistic annotation categories existed, and virtually none had been developed for annotation scheme design.

In the early 1990s, annotated corpora were typically regarded as stand-alone resources that would be used in isolation and not combined with other resources containing other annotation types. The primary motivation to standardize formats or categories during this period was to make them re-usable with different processing tools, or for the purpose of evaluation. A few years later, researchers in the U.S. began to pay more attention to harmonization of annotation practices in organized projects such as the Discourse Resource Initiative [8], and within programs such as DARPA's Message Understanding Conferences (MUCs) [20] and the Automatic Content Extraction (ACE) Program [14], which developed annotation guidelines for phenomena such as basic named entity classes and coreference to facilitate evaluation–some of which served as *de facto* standards for several years following. In the next decade, the need for standards gained considerably more at-

[2] http://nlp.shef.ac.uk/parole/parole.html

[3] A few projects relied on manual annotation alone [31, 45, 33], partial "spot-checking" of automatically-generated annotations (e.g., the British National Corpus), or even combinations of several automatic annotators [41].

[4] http://www.MTurk.com

[5] See Part I, Chapter 3 in this volume for an overview.

[6] http://www.ilc.cnr.it/EAGLES/browse.html

[7] www.ilc.cnr.it/EAGLES/annotate/annotate.html

tention, as annotated data was more and more widely available and the obstacles to reuse–namely, lack of commonality of formats and schemes–became painfully apparent. Chapter IV provides a brief history of standardization efforts and surveys the standards for both linguistic annotation content and representation currently in use within the community.

Tools to support linguistic annotation proliferated when large-scale annotation projects began to be undertaken in the late 1980s and early 90s. For the most part, these early tools were developed in-house and geared toward a specific annotation task. Starting in the mid-1990s, a spate of general purpose annotation tools (typically referred to as annotation "architectures" or "workbenches") became available, including but not limited to the General Architecture for Text Engineering (GATE) [10], the Alembic Workbench [11], the Architecture and Tools for Linguistic Analysis Systems (ATLAS) [3], the Callisto annotation tool [12], the MATE (Multilevel Annotation Tools Engineering) workbench [30], and its successor NITE (Natural Interactivity Tools Engineering) [2]. The evolution of these tools is tightly coupled with standardization efforts for physical representation of linguistically-annotated data due to their implementation of several competing physical formats developed during this period. Many of these architectures and workbenches have since faded into history; the notable exception is GATE, which provides for manual annotation as well as (and primarily) pipelining annotation tools whose output can then be manually edited. The Unstructured Information Management Applications (UIMA) [16] is a more recent, widely-used framework that provides similar capabilities and implements (yet another) representation format; these two major frameworks are described in Wilcock (Chapter V.b).

Several platforms devoted specifically to speech annotation were also developed in the 90s, providing for time-aligned annotation of audio signals with orthographic transcriptions and linguistic phenomena such as such as prosody and phonetics. Similar tools have been developed, especially over the past two decades, for annotating video signals for gesture, sign language transcription, etc., some of which are extensions of tools originally designed for speech annotation. Cassidy and Schmidt (Chapter V.c) provide a comprehensive inventory of state-of-the-art tools for multimodal annotation and the range of standard means to represent them.

It has become increasingly common to establish annotation projects where annotators are located at different sites around the world, and who access and annotate data over a relatively long time period through a web-based interface. Tool support for this kind of activity is relatively new; it requires the means to manage versions of the annotated data as they are modified, possibly simultaneously, by multiple users, etc. Bieman *et al.* (Chapter V.d) outline requirements for web-based annotation tools and review a variety of existing tools. More generally, Finlayson and Erjavec (Chapter V.a) outline the process of creating end-to-end linguistic annotations and assess the requirements for annotation tool design, regardless of purpose or procedure.

Manual annotation projects in the early 1990s attempted to measure consistency and agreement among annotators, but there were few established practices in the field. The Penn Treebank project gave annotators 10% overlapped material in order to evaluate consistency of predicate-argument structure added to the Treebank in the

mid-90s [36]. Annotation efforts involving more subjective phenomena computed agreement using a variety of methods (e.g., [34, 47, 42]) until Carletta's seminal paper [5] proposed borrowing the Kappa coefficient of agreement from the field of content analysis [46]. From the mid-90s onward there was a dramatic increase in reports of inter-annotator agreement (IAA) for linguistic annotation across a broad range of phenomena (e.g., word sense disambiguation [44, 39], translation equivalents [38], discourse parsing and labeling[35]), and since then, Kappa has served as one of the primary "go-to" statistics for measuring IAA in the field along with a handful of others (e.g., Krippendorf's Alpha). Recent work has suggested a variety of alternatives to standard measures [1]; see Artstein (Chapter V.g.) in this volume for a comprehensive overview. The following chapter, by Takenabu (Chapter V.h), describes experimentation with a novel methodology for analyzing annotator agreement by collecting data on annotation tool operation and annotator eye gaze and mapping the behavior to agreement levels.

The initiation of the Language Resources and Evaluation Conference (LREC) in 1998 and the subsequent creation of a journal of the same name (*Language Resources and Evaluation*[8], Springer) had broad impact on both the number of linguistic annotation efforts and the perceived validity of annotated resource creation as a worthy scholarly activity, by providing a venue for presentation and discussion of annotation practices and results. In 2007, the field was further legitimized when the Association for Computational Linguistics established a Special Interest Group for Linguistic Annotation (SIGANN)[9], which has held an annual workshop (Linguistic Annotation Workshop: the LAW) since then. As a result, methods for the design and application of linguistic annotation schemes have become increasingly formalized over the past fifteen years, leading to a set of practices that have been referred to as "annotation science" [26, 25, 22] as well as formal methods for annotation scheme design [4, 28] and sophisticated frameworks for physical representation [23, 21]. Pustejovsky *et al.* (Chapter II) is concerned with the criteria and methodology for annotation scheme design–i.e., definition of labels and features describing linguistic phenomena and the relationships among them that comprise the annotation scheme. Ide *et al.*(Chapter III) examine the other side of scheme design: identification of a physical, machine-readable format that can capture the required information and is easily and flexibly processable. Each of these aspects of annotation scheme design has undergone extensive development over the past fifteen years; these two chapters discuss in detail these developments together with the current state-of-the-art and "best practices" in the field today.

In the 1980s, linguistic annotation was usually motivated by the desire to study a given linguistic phenomenon in large bodies of data, and annotation schemes typically directly reflected a specific linguistic theory. As the need for reliable automatic annotation for larger and larger bodies of data increased in the early 90s, there sometimes arose a tension between the requirements for accurate automatic annotation and a comprehensive linguistic accounting that could contribute to validation and re-

[8] http://link.springer.com/journal/10579

[9] http://www.cs.vassar.edu/~sigann

finement of the underlying theory. An early example is the Penn Treebank project's reduction and modification of the part-of-speech tagset developed for the Brown Corpus, in order to obtain more accurate results from automatic taggers and parsers. In the following decades, machine learning arose as the central methodology for NLP; therefore, some annotation projects began to design schemes incrementally, relying on iterative training and re-training of learning algorithms to develop annotation categories and features in order to best tune the scheme to the learning task (see, for example, [43])–in a sense shifting 180 degrees from *a priori* scheme design based on theory to *a posteriori* scheme development based on data, and potentially limited by constraints on feature identification. Despite the increasing prevalence of this approach, there has been little discussion of the impact and value of iterative scheme development in the service of machine learning.

Two chapters in Part I, Section VI of this volume ("Using annotations") are concerned with the role of linguistic theory in annotation, most directly de Marneffe and Potts (Chapter VI.d), who take issue with the common wisdom that annotated corpora are primarily useful for building computational models and contribute little or nothing to linguistic inquiry. They argue that all linguistic annotations are a product of theoretical assumptions and intuitions which, once identified, provide a sound basis for developing and testing linguistic theories and outline some strategies for doing so. Flickinger *et al.* (Chapter VI.d) discuss the development of complex syntactico-semantic annotations grounded in the theoretical framework of Head-Driven Phrase Structure Grammar (HPSG), via a novel method of incremental improvement in which all manual effort, including annotation design and disambiguation, is encoded in such a way that its value is preserved and enhanced over time, and ultimately can be reused by the machine.

Finally, Part I Section VI contains two chapters covering major activities in which linguistically-annotated data play a central role. Rumshisky and Stubbs (Chapter VI.a) discuss the use of these data in machine learning, and describe how data annotated at multiple linguistic levels are leveraged to generate sophisticated language models for NLP. Gries and Berez (Chapter VI.c) overview the use of linguistic annotations in corpus linguistics, providing a survey of annotation types of interest to this field and the format and contents of resources commonly exploited by corpus linguists.

4 Part II: Case Studies

The primary goal for including an extensive set of annotation case studies in this volume is to provide guidance for future annotation efforts and demarcate current practice, thereby contributing to the continued evolution of best practices for the field. To address this goal, the contributing authors were provided with a set of guidelines and encouraged to be as candid as possible in describing their project, its methodology, outcomes, and "lessons learned", which is shown in Figure 1.

Guidelines for Case Studies

Each case study should provide an overview of the annotation project, its purpose and results, and place it in the historical context of similar projects. The description should address the issues discussed in Part I, including those in the outline below. If one of these issues is of particular relevance or importance in your project, it can serve as a focus for the chapter (but please try to address the range of issues to the extent that they apply). If there are issues not addressed in the book or outline that seem relevant, feel free to address them as well.

The case study should not be just a tech report, but should provide background and motivation that will be helpful to others who undertake annotation projects. Be honest! Annotation projects often have to cut corners and readers want to learn from past experiences. Your opinions on what did or did not work well will provide valuable information for future projects.

1. Annotation scheme:
    a. What is the underlying theory?
    b. How were the features included in the scheme chosen?
    c. What was the process of development (iteration over annotation exercises, etc.)
    d. Has the potential use of the annotations informed development of the annotation scheme?
    e. Has development of the scheme informed the development of linguistic theories or knowledge?
2. Physical representation:
    a. How is the annotation represented?
    b. Why was this representation chosen?
    c. What are the advantages/disadvantages of this representation that may have come to light through its use?
    d. What software or system was used to generate the annotated data?
3. Annotation Process:
    a. Was the annotation done manually, automatically, or via some combination of the two?
    b. Manual annotation:
        i. How many annotators were involved, what was their background, etc.
        ii. What annotation environment was used (e.g., GATE)?
        iii. What was the exact process by which annotations were done? Multiple steps, multiple annotators, etc.
        iv. Was inter-annotator agreement computed and if so, by what method and what were the results?
    c. Automatic annotation:
        i. What software was used to generate the annotations?
        ii. How well does this software generally perform? Did it perform better or worse on your data?
4. Evaluation/Quality control: By what method(s) was the quality of the annotations evaluated?
5. Usage:
    a. By what means and under what conditions is the data available to users?
    b. What were the expected usages of the annotated data? What are the actual uses of the data, if different?
    c. If your corpus has been used as training data for a machine learning algorithm, what was the task? How much did the linguistic annotation contribute to the performance of classification (or other learning tasks), above and beyond *n*-gram features already present in the corpus?

Fig. 1 Guidelines for case study authors

The case studies in this volume describe major annotation projects over a broad range of phenomena at different linguistic levels for text and speech as well as multimodal data. While it was not possible to obtain a case study for every annotation project that might deserve inclusion, the thirty-nine exemplars provide a comprehensive overview of the state-of-the-art in the field. Collectively, they cover projects that annotate data across two or more genres as well as data from specialized domains, in particular, chemical, biological, and medical data (see the four case studies in section VI). The majority annotate data in a single language–primarily English, but also Czech (II.a; III.d.i), Chinese (II.c), French (IV.b.iii; V.c), German (II.b), Arabic (III.b.i), Turkish (IV.b.ii), and Japanese (IV.a.ii; IV.c.i), but several annotate over multiple languages: e.g., English/Chinese/Arabic (I.d), Hindi/Urdu (II.d), six Nordic dialects (V.b), and the sixteen primarily Eastern European languages included in MULTEXT-EAST (I.a)–the last of which is the only project including parallel aligned data.

The project descriptions include several that focus on a specific linguistic phenomenon (e.g., metaphor, word senses, sentiment, dialogue acts, factivity, temporal and spatial information, textual entailment, etc.), but also include a large number that annotate multiple linguistic layers. The corpora described in Chapters 16-19 were all designed to cover a range of phenomena at different linguistic levels, and, although purportedly dedicated to syntax or discourse, the various treebanks described in Chapters 20-23 invariably include multiple layers with related syntactico-semantic information. The full list of annotation types covered in Part II is shown in Figure 2.

Topic	Chapter numbers
General corpora	16, 17, 18, 19
Treebanks	20, 21, 22, 23
Sense tagging	24, 25
Semantic roles	25, 26, 27
Opinion, sentiment, subjectivity	28, 29
Named entities	30, 31, 32
Factivity	33
Time and event annotation	34, 35
Spatial phenomena	36, 37
Metaphor	38, 39
Textual entailment	40, 41
Coreference	42, 43
Discourse structure	44, 45, 46
Dialogue Acts	47
Speech	48, 49, 50
Biomedical annotations	51, 52, 53, 54

Fig. 2 Summary of topics and case studies in Part II.

One of the common themes in case studies where different annotation types are layered is the difficulty of combining annotations that are produced using different tools and usually represented in different formats. Several different approaches were adopted to solve the problem. The Corpus of Interactional Data (CID, V.c) was

faced with the challenge of harmonizing multiple layers across modalities, including prosody, phonetics, morphology, syntax, lexical semantics, gestures, attitudes, etc. Their solution was to utilize an abstract model of typed feature structures for all annotation types, which enabled representing the different layers and the relations among them homogeneously, thereby facilitating search over the various types of information. Similarly, AusTalk (V.a) represents annotations of both audio, video, and transcriptions in the Resource Description Language (RDF). OntoNotes (I.d) translates all annotations into the relational database model; the project faced additional harmonization problems due to the dependence of annotation layers on the tokenization and syntactic structure of its three languages' treebanks, which were undergoing constant modification at the time (see I.d, section 4.2). MASC's (I.c) original annotations and all contributed annotations are represented in the Linguistic Annotation Framework's graph-based format (GrAF), in order to enable them to be merged together for the study of inter-layer phenomena. Interestingly, all of these representations are based on the same underlying abstract model, attesting to its universality for representing linguistic annotations (see Part I, Chapter III for a discussion). Other projects have taken the opposite approach and represent multiple annotation layers in different formats. The Hindi/Urdu treebank (II.d) contains three layers of annotation: dependency structure (DS), PropBank-style annotation for predicate-argument structure, and phrase-structure annotation, each with its own framework and annotation scheme. Layers of annotation in CRAFT (VI.b) are also represented in different formats, including the Penn Treebank bracketed format and the Knowtator format plus several alternative representations.

The case studies vary in their focus on particular aspects suggested in the case study guidelines. Some focus almost exclusively on the design and content of the applied annotation scheme and its rationale (e.g., for sense annotation (VerbNet, IIIa.ii), semantic roles (PropBank, III,b,i), treebanks (Sinica Treebank, II.c), clinical text (VI.a), text entailment (RTE, III.i.ii), metaphor (CMT, III.h.ii), dialogue acts (NICT, IV.c.i) , Japanese coreference (NAIST, IV.a.ii), biomedical data (GENIA, VI.c), spatial information (ISO-Space, III.g.I, and SRL, III.g.ii), time and event annotation (ISO-TimeML, III.f.i). The case studies for ANNODIS (IV.b.iii) and CMT (III.h.ii) spend considerable time describing the theory upon which the annotation scheme is based; these are the only two case studies that are deeply bound to a particular underlying theory. The Nordic dialogue case study (V.b) focuses almost entirely on issues of transcription, which are also covered in some detail in the AusTalk (V.a) and CID (V.c) chapters.

The studies reveal some interesting facts about the annotation tools that are used in actual practice. A few projects rely on a single, general-purpose platform, including GATE [9] (MASC, I.c; crowdsourced named entities, III.d.ii; and MPQA, III.c.i) and NITE [2] (GBM, I.b). A small number of projects performing ontology-based annotations use Knowtator [40] (JDPA, III.c.ii; CRAFT, VI.b; clinical texts, VI.a). Other projects use a suite of available tools (CID, V.c; Hindi/Urdu treebank, II.d; AusTalk, V.a; clinical texts, VI.a), some changing tools in mid-project to accommodate unmet needs (e.g., Ita-TimeBank, III.f.ii). The Czech Named Entity Corpus project (III.d.i) simply uses a text editor, and the VU metaphor, project (III.h.i)

uses the Oxygen XML editor[10]. However, a surprisingly large number of projects developed their own annotation tools to suit project needs, in some cases after experimentation with or extended use of existing tools; these projects include at least the following: TIGER (II.b), Prague Dependency Treebank (II.a), Chemical named entities (III.d.iii), GMB (I.b), FactBank (III.e.i), FrameNet (III.b.ii), NICT (IV.c.i), PropBank(III.b.i), TimeNL/TimeBank (III.f.i), and clinical text (VI.a). Finlayson and Erjavec (Part I, Chapter V.a) take the tendency for annotation projects to build from scratch the "right" annotation tool as a starting point and survey the functionality requirements for annotation tools, in order to provide a basis for identifying core and extension capabilities of an "all-purpose" annotation tool or, at least, determining why such a tool is not feasible.

Given the diversity of annotation tools used in the projects described in this volume, it is not surprising that the annotated data they produce are represented in a wide variety of physical formats. The vast majority of projects publish their annotated data in some flavor of XML, which is good news in terms of syntactic consistency, since, provided with the accompanying DTD or schema, the data can be read by any XML-aware tool, but to meaningfully process the data, the software must have some built-in knowledge of what to do with an element or attribute with a given name. Some of the XML formats referenced in the case studies serve as "meta-formats", in that they utilize XML elements to structure information rather than simply name it–for example, LAF/GrAF, TIGER-XML, and the CID project's feature structure-based format use XML elements to represent structural information (e.g., node, edge, terminal, constituents, etc.), while other XML-based formats identify annotation objects with XML element names. Other formats include tab-separated-values (FactBank, III.e.i), a column-based format (NEGRA, II.b), and TEI P5 (Multext-East, I.a; GENIA, VI.c; VU, III.h.i). Several of the annotated resources use a standoff representation, including Phrase Detectives (IV.a.i), MASC (I.c), FrameNet (III.b.ii), GMB (I.b), MPQA (III.c.i), Crowdsourcing Named Entities (III.d.ii), FactBank (III.e.i), JDPA (III.c.ii), CRAFT (VI.b), CID (V.c), clinical texts (VI.a), (Ita-TimeBank (III.f.ii), and PropBank(III.b.i). None of the case studies report on representing annotations as linked data (see Part I, Chapter III, Section 5.2) although AusTalk's (V.a) use of RDF obviously allows for that option, and MASC (I.c) has been rendered in linked format and included in the Linked Linguistic Open Data cloud[11].

The case studies describe annotation efforts that are entirely manual (e.g., FATE, III.i.i; FrameNet, III.b.ii) as well as a large number of projects in which automatically-produced annotations are hand-validated (e.g., MASC, I.c; RTE, III.i.ii; German Treebanks, II.b). Some projects do both for different phenomena as necessary (e.g., CID, V.c). The Hindi/Urdu Treebank project (II.d) manually annotated its dependency and semantic role layers, and then generated a phrase-structure layer automatically from the other two. The case studies also report on several emerging approaches to manual annotation/validation, including pair annotation (Turkish

[10] http://oxygenxml.com

[11] http://linguistic-lod.org/llod-cloud

Discourse Bank, Iv.b.ii), crowdsourcing (Crowdsourcing Named Entities, III.d.ii; MASC Sentence Corpus, III.a.i; RTE, III.i.ii), and games-with-a-purpose (Phrase Detectives, IV.a.i; GMB, I.b). i2b2 (VI.d) describes an in-depth comparison of serial annotation (annotation by annotators in succession) and parallel annotation (annotation by multiple annotators at once), which is also discussed in the CRAFT case study (VI.b). Most of the case studies provide detailed information on computing inter-annotator agreement as well.

It is interesting to note that the majority of the resources described in the thirty-nine case studies are either freely available or available under liberal licenses or agreements (e.g., restricted to research use). This is in contrast to the situation two decades ago, when manually annotated or validated language resources were often costly to obtain. This shift in community practice, together with the development of increasingly compatible annotation schemes and formats, means that high-quality annotated resources are now much more readily available to researchers throughout the world.

5 Conclusion

The past four decades have seen a great deal of evolution in strategies and "best practices" (*de facto* or otherwise) for linguistic annotation, spurred in particular by the need for gold standard data to train machine learning algorithms. Problematically, annotation practices and scheme design were relatively *ad hoc* when activity in the field stepped up in the 90s, and so development of more systematic and principled approaches has been to some extent hampered by the need to accommodate large amounts of legacy data, software, and the use of various *de facto* standards that are often inappropriate for any but the phenomenon for which they were designed. To this day, annotation efforts are plagued by the lack of something as basic as standardized tokenization procedures. Nonetheless, the past fifteen years have seen steady progress and convergence in harmonizing linguistic annotation practices and the resources that continue to be created, even if actual practice still falls short of our understanding of the science of linguistic annotation. This therefore seems to be an appropriate point for a volume on the topic that brings together the community's collective wisdom and experience, in order to lay the groundwork for further progress.

The primary target readership for this volume is the community of scholars and researchers who create, use, and distribute linguistically annotated resources. The volume should also be useful for students in undergraduate and graduate courses that create and/or use these data, especially when projects demand that students annotate data of their own for analysis. Finally, it may provide insight for those studying machine learning techniques that rely on gold standard annotations.

References

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008)
2. Bernsen, N.O., Dybkjær, L., Kolodnytsky, M.: The NITE Workbench. A Tool for Annotation of Natural Interactivity and Multimodal Data. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain (2002). URL http://www.lrec-conf.org/proceedings/lrec2002/pdf/214.pdf. ACL Anthology Identifier: L02-1214
3. Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., Liberman, M.: Atlas: A flexible and extensible architecture for linguistic annotation. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000). European Language Resources Association (ELRA), Athens, Greece (2000)
4. Bunt, H.: A methodology for designing semantic annotation languages exploiting semantic-syntactic isomorphisms. In: Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL2010), pp. 29–46. City University of Hong Kong, Hong Kong SAR (2010)
5. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics **22**(2), 249–254 (1996)
6. Church, K.W.: A stochastic parts program and noun phrase parser for unrestricted text. In: Proceedings of the Second Conference on Applied Natural Language Processing, ANLC '88, pp. 136–143. Association for Computational Linguistics, Stroudsburg, PA, USA (1988). DOI 10.3115/974235.974260. URL http://dx.doi.org/10.3115/974235.974260
7. Clear, J.H.: The british national corpus. In: G.P. Landow, P. Delany (eds.) The Digital Word, pp. 163–187. MIT Press, Cambridge, MA, USA (1993)
8. Core, M., Ishizaki, M., Moore, J., Nakatani, C., Reithinger, N., Traum, D., Tutiya, S.: The report of the third workshop of the discourse resource initiative. Tech. rep., Chiba University and Kazusa Academia Hall (1998)
9. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust nlp tools and applications. In: Proceedings of ACL'02 (2002)
10. Cunningham, H., Wilks, Y., Gaizauskas, R.: Software Infrastructure for Language Engineering. In: Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition. Brighton, U.K. (1996)
11. Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., Vilain, M.: Mixed-initiative development of language processing systems. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 348–355. Association for Computational Linguistics, Washington, DC, USA (1997)
12. Day, D.S., McHenry, C., Kozierok, R., Riek, L.: Callisto: A configurable annotation workbench. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004). European Language Resources Association (2004)
13. DeRose, S.J.: Grammatical category disambiguation by statistical optimization. Computational Linguistics **14**(1), 31–39 (1988)
14. Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S., Weischedel, R.M.: The automatic content extraction (ace) program - tasks, data, and evaluation. In: Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2004. European Language Resources Association (2004)
15. Erjaveç, T., Ide, N.: The Multext-East Corpus. In: Proceedings of First International Conference on Language Resources and Evaluation, pp. 971–974 (1998)
16. Ferrucci, D., Lally, A.: Uima: An architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering **10**(3-4), 327–348 (2004)
17. Garside, R.: The CLAWS word-tagging system. In: R. Garside, G. Sampson, G. Leech (eds.) The Computational Analysis of English: : A Corpus-Based Approach. Longman (1987). URL http://www.researchgate.net/publication/230876041_The_CLAWS_word-tagging_system

18. Garside, R., Leech, G., Sampson, G.: The computational analysis of english: a corpus-based approach. Longman (1987)
19. Greene, B.B., Rubin, G.M.: Automatic Grammatical Tagging of English. Department of Linguistics, Brown University (1971)
20. Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: Proceedings of the 16th Conference on Computational Linguistics - Volume 1, COLING '96, pp. 466–471. Association for Computational Linguistics, Stroudsburg, PA, USA (1996)
21. Hellmann, S., Lehmann, J., Auer, S., Nitzschke, M.: Nif combinator: Combining nlp tool output. In: 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW2012) (2012)
22. Hovy, E., Lavid, J.: Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. International Journal of Translation Studies **22**(2) (2010)
23. Ide, I., Suderman, K.: The linguistic annotation framework: A standard for annotation interchange and merging. Language Resources and Evaluation **48**(3), 395–418 (2014)
24. Ide, N.: Corpus encoding standard: Sgml guidelines for encoding linguistic corpora. In: Proceedings of the First International Language Resources and Evaluation Conference (LREC 1998), pp. 463–70. European Language Resources Association (ELRA) (1998)
25. Ide, N.: Annotation science: From theory to practice and use. In: G. Rehm, A. Witt, L. Lemnitzer (eds.) Data Structures for Linguistics Resources and Applications. Gunter Narr Verlag, Tübingen, Germany (2007)
26. Ide, N., Atwell, E. (eds.): Annotation Science: State of the Art in Enhancing Automatic Linguistic Annotation: Proceedings of the Workshop. European Language Resources Association (2006). URL http://www.lrec-conf.org/proceedings/lrec2006/
27. Ide, N., Bonhomme, P., Romary, L.: Xces: An xml-based encoding standard for linguistic corpora. In: Proceedings of the Second Language Resources and Evaluation Conference (LREC 2000). European Language Resources Association (ELRA), Athens, Greece (2000)
28. Ide, N., Bunt, H.: Anatomy of annotation schemes: Mapping to graf. In: Proceedings of the Fourth Linguistic Annotation Workshop, LAW IV '10, pp. 247–255. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
29. Ide, N., Véronis, J.: MULTEXT: multilingual text tools and corpora. In: Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), vol. I, pp. 588–592. Kyoto, Japan (1994)
30. Isard, A., Mller, M.B., McKelvie, D., Mengel, A.: The mate workbench - a tool for annotating xml corpora. In: Proceedings of Recherche d'Informations Assiste par Ordinateur (RIAO'2000). Paris (2000)
31. Jäborg, J.: Introduction to "This is Watson". Göteborg University, Institute för språkvetenskaplig databehandling (1986)
32. Kuĉera, H., Francis, W.N.: Computational Analysis of Present-Day American English. Brown University Press, Providence, RI, USA (1967)
33. Landes, S., Leacock, C., Tengi, R.I.: Building semantic concordances. In: C. Fellbaum (ed.) WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, Massachusetts (1998)
34. Litman, D., Hirschberg, J.: Disambiguating cue phrases in text and speech. In: Proceedings of the 13th Conference on Computational Linguistics - Volume 2, COLING '90, pp. 251–256. Association for Computational Linguistics, Stroudsburg, PA, USA (1990)
35. Marcu, D., Amorrortu, E., Romera, M.: Experiments in constructing a corpus of discourse trees. In: Towards Standards and Tools for Discourse Tagging, pp. 48–57 (1999)
36. Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The penn treebank: Annotating predicate argument structure. In: Proceedings of the Workshop on Human Language Technology, pp. 114–119. Association for Computational Linguistics, Stroudsburg, PA, USA (1994)
37. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of english: The penn treebank. Computational Linguistics **19**(2), 313–330 (1993)
38. Melamed, I.D.: Manual annotation of translational equivalence: The blinker project. CoRR **cmp-lg/9805005** (1998)

39. Ng, H.T., Lim, C.Y., Foo, S.K.: A case study on inter-annotator agreement for word sense disambiguation. In: SIGLEX99: Standardizing Lexical Resources, pp. 351–14 (1999)
40. Ogren, P.V.: Knowtator: A protÉgÉ plug-in for annotated corpus construction. In: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume: Demonstrations, pp. 273–275. Association for Computational Linguistics, Stroudsburg, PA, USA (2006)
41. Paroubek, P.: Language resources as by-product of evaluation: The multitag example. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000). European Language Resources Association (ELRA), Athens, Greece (2000)
42. Passonneau, R.J., Litman, D.J.: Intention-based segmentation: Human reliability and correlation with linguistic cues. In: Proceedings of the 31st Annual Meeting on Association for Computational Linguistics, ACL '93, pp. 148–155. Association for Computational Linguistics, Stroudsburg, PA, USA (1993)
43. Pustejovsky, J., Stubbs, A.: Natural language annotation for machine learning. O'Reilly Media, Sebastopol, CA (2013)
44. Resnik, P.: Disambiguating noun groupings with respect to wordnet senses. Proceedings of the 3rd Workshop on Very Large Corpora (1995)
45. Sampson, G.: English for the Computer: The SUSANNE Corpus and Analytic Scheme. Clarendon Press (1995)
46. Siegel, S., Castellan, N.: Nonparametric statistics for the behavioral sciences, second edn. McGraw–Hill, Inc. (1988)
47. Silverman, K.E.A., Beckman, M.E., Pitrelli, J.F., Ostendorf, M., Wightman, C.W., Price, P., Pierrehumbert, J.B., Hirschberg, J.: Tobi: a standard for labeling english prosody. In: International Conference on Spoken Language Processing. ISCA (1992)