Management, Sustainability, and Interoperability of Linguistic Annotations

Nancy Ide
Department of Computer Science
Vassar College

1 Introduction

In recent years, there has been a noticeable upswing in linguistic annotation activity, which has expanded to cover a wide variety of linguistic phenomena. At the same time, the number and size of linguistically annotated language resources has increased dramatically, together 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. All of this has created a need to manage and sustain these resources as well as to find ways to enable them to be repeatedly reused and merged with other resources.

2 What is Linguistic Annotation?

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, co-reference 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.

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.

A linguistic annotation scheme is comprised of two main parts: the scheme's *semantics*, which specify the categories and features that label and provide descriptive information about the data with which they are associated, and the scheme's *representation*, which is the physical format in which the annotation information is represented for consumption by software (and in some cases, humans as well). Historically, designers of linguistic annotation schemes have focused on determining the appropriate categories and features to describe the phenomenon in question and paid less attention to the eventual physical representation of the annotation information, with possibly unintended results when constraints imposed by the physical representation affect choices for the conceptual content of an annotation scheme. In recent years, the need to compare and combine annotations as well as use them in software environments for which they may have not been originally designed has increased, leading to the awareness that a conceptual scheme may be represented in any of a variety of different physical formats and/or transduced from one to the other.

Both the syntax and semantics of an annotation scheme involve choices that are, to some extent, arbitrary, but which have ramifications for their usability. For the physical format, the most significant choice is whether to insert the annotation information into the data itself or represented in *standoff* form, that is, provided in a separate document with links to the positions in the original data to which each annotation applies.

3 History

In the mid-twentieth century, linguistics was practiced primarily as a descriptive field, studying structural properties within a language and typological variations between languages. This work resulted in fairly sophisticated models of the different informational components comprising linguistic utterances. As in the other social sciences, the collection and analysis of data was also subjected to quantitative techniques from statistics, and in the 1940s, linguists such as Bloomfield and others were starting to think that language could be explained in probabilistic and behaviorist terms. At the same time in the related and emerging field of NLP, Warren Weaver suggested using

computers to translate documents between natural human languages, and in 1949 produced a memorandum entitled “Translation" (Weaver 1955) outlining a series of methods for that task. Empirical and statistical remained popular throughout the 1950s, and Shannon's information-theoretic view to language analysis provided a solid quantitative approach for modeling qualitative descriptions of language. However, datasets were generally so small that it was not possible to extract statistically significant patterns to support probabilistic approaches, and as a result, linguistically annotated corpora did not play a major role in the first years of NLP (1950s–60s).

In the 1960s, there was a general shift in the social sciences, particularly in the United States, from data-oriented descriptions of human behavior to introspective modeling of cognitive functions. As part of this new attitude towards human activity, the U.S. linguist Noam Chomsky focused on both a formal methodology and a theory of linguistics that not only ignored quantitative language data, but also claimed that it was misleading for formulating models of language behavior. Chomsky's view was influential in the United States throughout the next two decades, largely because the formal approach enabled the development of extremely sophisticated rule-based language models using mostly introspective (or self-generated) data, providing an attractive alternative to creating statistical language models on the basis of relatively small data sets of linguistic utterances from the existing corpora in the field. In NLP, the flourishing field of Artificial Intelligence (AI) began to attack the problem of language understanding and, in the spirit of the times, abandoned empirical methods and grounded language processing system design in formal theories of human language understanding, which they attempted to model. IBM's championing of statistical methods for speech processing in the 70s and 80s was one of the few efforts that went against this trend during that era. Reasonably large linguistically annotated resources were relatively rare; a well-known exception is the one million-word Brown Corpus of Standard American English (Kuĉera and Francis 1967). In the 1970s, the Brown Corpus was the object of what is arguably the first modern linguistic annotation project, which added part-of-speech annotations.[1] Like the Brown Corpus, corpora developed in the 70s and 80s were typically annotated only for part-of-speech, because 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 lin-

[1] It is interesting to note that the Brown Corpus annotation project fostered the development of increasingly accurate automatic methods for part-of-speech tagging in order to avoid the painstaking work of manual validation.

guistic phenomena, such as syntax.[2]

All of this changed in the mid- to late-1980s, when large-scale language data resources began to become available. This led to a proliferation of linguistic annotation projects, most still focused on part-of-speech (or richer morpho-syntactic) annotations, and spearheaded the re-introduction 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 (Garside 1987). Building on this work, the Penn Treebank project (Marcus, Marcinkiewicz, and Santorini 1993) 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 (Marcus et al. 1994). 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 (Clear 1993), released in 1994; corpora produced by the MULTEXT project (1993-96) (Ide and Véronis 1994) and its follow-on, MULTEXT-EAST (1994-97) (Erjaveç and Ide 1998), which provided parallel aligned corpora in a dozen Western and Eastern languages annotated for part-of-speech; and the PAROLE and SIMPLE corpora[3], which included part-of-speech tagged data in fourteen European languages. Following these efforts, syntactic treebanks for a wide variety of languages (e.g., Swedish, Czech, Chinese, French, German, Spanish, Turkish, Italian) proliferated over the next decade, together with corpora annotated for other phenomena, e.g., word sense annotations (SemCor (Landes, Leacock, and Tengi 1998)), which similarly engendered the development of comparably-annotated corpora in other languages (Bentivogli, Forner, and Pianta 2004; Lupu, Trandabăţ, and Husarciuc 2005; Bond et al. 2012).

During this period, linguistic annotation was often 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. Designers of linguistic annotation schemes focused on determining the appropriate categories and features to describe the phenomenon in question, and paid less attention to the eventual *physical representation* for the annotations in the resource. Insofar as physical format was considered, the criteria for determining them was invariably the ease of processing by soft-

[2]The earliest automatic part-of-speech taggers include Greene and Rubin's TAGGIT (Greene and Rubin 1971), Garside's CLAWS (Garside 1987), DeRose's VOLSUNGA (DeRose 1988), and Church's PARTS (Church 1988).

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

ware that would use the output. For example, early formats for phenomena such as part of speech often output one word per line, separated from its part of speech (POS) tag by a special character such as an underscore or slash (DeRose 1988; Church 1988) (see Figure 09.01). Syntactic parsers producing constituency analyses typically used what has come to be known as the “Penn Treebank format", which brackets and nests constituents with parentheses, LISP-style (Marcus, Marcinkiewicz, and Santorini 1993; Charniak 2000; Collins 2003).

[figure 09.01 about here]

Dependency parsers often used a line-based format that provides the syntactic function and its arguments in specified fields. Interestingly, these early formats for POS tagger and parser output have remained in use, with very little variation, up to the present day, primarily in the output of POS taggers; see for example, the Stanford taggers and parsers for multiple languages[4], TreeTagger[5], and TnT[6]. Such formats rely heavily on white space and line breaks, together with occasional special characters, to delineate elements of the analysis (e.g., individual tokens and part of speech tags). As a result, software intended to use these formats as input must be programmed to understand the meaning of these separators, together with the nature of the information in each field.

The separation between conceptual content and physical representation has not always been taken into account when schemes are designed, with possibly unintended results; for example, a representation format may impose limits on the complexity of the information that can be included or force the conflation of information into cryptic labels that may be impossible to later disentangle. In recent years, the need to compare and combine annotations as well as use them in software environments for which they may have not been originally designed has increased, leading to the awareness that a conceptual scheme may be represented in any of a variety of different physical formats and/or transduced from one to the other. Experience with annotated data that is difficult to transduce or modify has engendered annotation “best practices" that dictate that annotation information be explicit (so it can be readily retrieved) and flexible (so other information can be substituted or added).

As the need for reliable automatic annotation for larger and larger bodies of data increased, there sometimes arose a tension between the requirements

[4]ttp://nlp.stanford.edu/software/tagger.stml

[5]ttp://www.cis.uni-muencen.de/ schmid/tools/TreeTagger/

[6]ttp://www.coli.uni-saarland.de/~torsten/tnt/

for accurate automatic annotation and a comprehensive linguistic accounting that could contribute to validation and refinement 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, (Pustejovsky and Stubbs 2012))–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.

4 The rise of standards

Over the past 30 years, generalized solutions for representing annotated language data–i.e., solutions that can apply to a wide range of annotation types and therefore allow for combining multiple layers and types of linguistic information–have been proposed.[7] The earliest format of note is the Standard Generalized Markup Language (SGML; ISO 8879:1986) (Iso 1986), which was introduced in 1986 to enable sharing of machine-readable documents, with no special emphasis on (or even concern for) linguistically-annotated data. Like its successor, the Extensible Markup Language (XML) (Bray et al. 2006), SGML defined a "meta-format" for marking up, or annotating, electronic documents consisting of rules for separating markup (tags) from data (by enclosing identifying names in angle brackets) and providing additional information in the form of attributes (features) on those tags.[8] SGML also specified a context-free language for defining tags and the valid structural relations among them (nesting, order, repetition, etc.) in an *SGML Document Type Definition* (DTD) that is used by SGML-aware software to validate the appropriate use of tags in a conforming document. XML replaced the DTD with the XML schema, which performs the same function

[7] Several initiatives have focused on reusability of language data from the late 1980s onward.

[8] Note that the Hypertext Markup Language (HTML) is an *application* of SGML/XML, in that it uses the SGML/XML meta-format to define specific tag names and document structure for use in creating web pages.

as well as some others.

The Text Encoding Initiative (TEI)[9] Guidelines, first published in 1992, defined a broad range of SGML (and later, XML) tags and accompanying DTDs for encoding language data. However, the TEI was from its beginnings intended primarily for humanities data and does not provide guidelines for representing many phenomena of interest for linguistic annotation. Therefore, in the mid-1990s, the EU EAGLES project[10] defined the Corpus Encoding Standard (CES) (Ide 1998), a customized application of the TEI providing a suite of SGML DTDs for encoding linguistic data and annotations, which was later instantiated in XML (XCES) (Ide, Bonhomme, and Romary 2000). In part as a result, SGML (and later, XML) began appearing in annotated language data in the mid-1990s, for example, in corpora developed in EU-funded projects such as PAROLE, data used in the US-DARPA Message Understanding Conferences (MUC) (Grishman and Sundheim 1995), and the TIPSTER annotation architecture (Grishman 1998) defined for the NIST Text Retrieval Conferences (TREC)[11], which included a CES-based SGML format for exporting output from information extraction tasks. SGML and XML were also adopted by major annotation frameworks developed during this period, such as GATE[12] and NITE[13], for import and export of data.

Although widely adopted, XML as an in-line format for representing linguistic annotations did not solve the reusability problem, for several reasons. First and foremost, XML requires that in-line tags are structured as a well-formed tree, thus disallowing annotations that form overlapping hierarchies and making connections between discontiguous portions of the data cumbersome. In addition, like all in-line formats, the insertion of of annotation information directly into the data imposes linguistic interpretations that may not be desired by other users. This includes segmental information–e.g., delineation of token boundaries in-line, whether by surrounding a string of characters with XML tags or separating it with white space, line breaks, or other special characters–as well as the inclusion of specific annotation labels and features. To solve this problem, in 1994 the notion of *stand-off annotation* was introduced in the CES[14], wherein annotations are maintained in

[9] w.tei-c.org/

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

[11] ttp://www-nlpir.nist.gov/related_projects/tipster/trec.tm

[12] http://gate.ac.uk

[13] http://groups.inf.ed.ac.uk/nxt/index.shtml

[14] Originally called "remote markup"–see ttp://www.cs.vassar.edu/CES/CES1-5.tml#ToCOview

separate documents and linked to appropriate regions of primary data, rather than interspersed in the primary data or otherwise modifying it to reflect the results of processing. This allows different annotations for the same phenomenon to co-exist, including variant segmentations (e.g. tokenizations) as well alternative analyses produced by different processors and/or using different annotation labels and features.

Annotation Graphs (AG) (Bird and Liberman 2001), introduced in 2001, are a standoff format that represents annotations as labels on edges of multiple independent graphs defined over text regions in a document. Because the model was developed primarily with speech data in mind, the regions are typically defined between points on a timeline, although this is not necessary. However, because each annotation type or layer is represented using a separate graph, the AG format is not well-suited to representing hierarchically-based phenomena such as syntactic constituency.[15]

Over the past decade, there has been an increasing convergence of practice for representing linguistic annotations in the field, with the aim of ensuring maximal reusability but also reflecting advances in our understanding of means to best structure and organize data, especially linked data intended for access and query over the web. In addition to the use of stand-off rather than in-line annotations, focus has shifted from identifying a single, universal format to defining an underlying data model for annotations that can enable trivial, one-to-one mappings among representation formats without loss of information. The most generalized implementation of this approach is the International Standards Organization (ISO) 24612 Linguistic Annotation Framework (LAF) (ISO 2012; Ide and Suderman 2014), which was developed over the past fifteen years to provide a comprehensive and general model for representing linguistic annotations. To accomplish this, LAF was designed to capture the general principles and practices of both existing and foreseen linguistic annotations, including annotations of all media types such as text, audio, video, image, etc., in order to allow for variation in annotation schemes while at the same time enabling comparison and evaluation, merging of different annotations, and development of common tools for creating and using annotated data.

LAF specifies a set of fundamental architectural principles, including the clear separation of primary data from annotations (i.e., standoff annotation); separation of annotation structure (i.e., physical format) and annotation content (the categories or labels used in an annotation scheme to describe

[15] An *ad hoc* mechanism to connect annotations on different graphs was later introduced into the AG model to accommodate hierarchical relations.

linguistic phenomena); and a requirement all annotation information be explicitly represented rather than building knowledge about the function of separators, position, etc. into processing software. It also defined an abstract data model for annotations, consisting of an acyclic di-graph decorated with feature structures, grounded in n-dimensional regions of primary data.

The LAF data model and architectural principles, which in large part simply brought together existing best practices from a variety of sources, significantly influenced subsequent development of models and strategies to render linguistic annotations maximally interoperable. As a result, most general-purpose physical formats developed over the past decade embody most if not all of LAF's principles. Formats to enable interoperability within large systems and frameworks have also followed many of the same principles and practices, for example, the Unstructured Information Management Architecture's (UIMA) (Ferrucci and Lally 2004) Common Analysis System (CAS), and the recently developed Language Applications Grid Interchange Format (LIF) (Verhagen et al. 2015), a JSON-LD-based format designed for interchange among language processing web services. The convergence of practice around the graph-based data model has led to the realization of increased compatibility of formats via mapping, and, as a result, transducers among formats are increasingly available that allow for the processing of annotated language resources by different tools and for different purposes (e.g., ANC2Go (Ide, Suderman, and Simms 2010), Pepper (Zipser and Romary 2010), and transducers available with DKPro[16] and the Language Applications (LAPPS) Grid[17]).

However, there is one widely-used format developed over the past two decades that does not follow LAF's principles. The desire for processing ease and readability fostered development of a simple column-based format for annotations for use in the Conference on Natural Language Learning (CoNLL) exercises. Most recently, a major project has developed a standard based on this format called CoNLL-U, the Universal Dependencies (UD) annotation format (Nivre et al. 2016). In this scheme, annotations are rooted in a fixed tokenization, itemized in a single column, and are not linked to primary data. Each column corresponds to a defined annotation type, indicating if the token in each row "begins" the annotation, is "inside" the annotation, or is "outside" the annotation.[18] Nested annotations, such as a constituency parse, are difficult to represent in this format without exploding

[16] http://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/

[17] http://lappsgrid.org

[18] The three possibilities are designated with "B", "I", and "O", respectively; the CoNLL format is often called the "BIO" format as a result.

the number of columns; to be fair, UD is intended primarily for dependency parses that do not present this problem. Alternative annotations of a given type cannot be represented easily because each column in the UD format has pre-defined content, and each row provides information for the token at its head. Other kinds of representation require even more gymnastics, if possible at all: for example, linking a given token such as the German “im" to its full form “in dem", which should be represented in two separate lines, thus disturbing the one-item-per-numbered-row scheme. Mapping UD or any similar column-based format to almost any other format is problematic, at best, thus hampering interoperability. However, the ease of processing and readability of this format have made these formats very popular, and they are not likely to be abandoned any time soon.

5 Interoperability as a focus

Over the past fifteen years, what was referred to as “reusability" in the 1990s came to be known as “interoperability". Over this period, the need for interoperability for linguistically-annotated resources became increasingly urgent, as more and more language data was being annotated for more than one type of linguistic phenomenon, and the need to use these annotations together was becoming more apparent. An experiment in the mid-2000s served to bring the need for annotation interoperability to the fore, especially in the US where it had been less a concern than in Europe: a project funded by the US National Security Agency called for annotation projects at labs around the US to annotate the same data (the 10,000 word Language Understanding (LU) corpus, or “Boyan 10K") for a wide variety of linguistic phenomena in order to study inter-level interactions. The annotations included syntax, semantic roles, opinion, committed belief, and others. Ultimately, it was determined that it was impossible to combine the annotations, due to differences in formats, labels for the same phenomena, conceptions of what is a relation and what is an object, and a loss of information implicit in the original representations when combining was attempted. The most insurmountable problem was a huge variation in tokenization practices, which are often minimally documented or not documented at all.

Beyond these difficulties, the definition of what it means for linguistic annotations to be interoperable is unclear, but obviously necessary in order assess the current state of interoperability in the field and measure our progress towards achieving interoperability in the future. What is needed is an *operational definition*, which identifies one or more specific observable

conditions or events that can be reliably measured, and where the results of the process are replicable.

Broadly speaking, interoperability can be defined as a measure of the degree to which diverse systems, organizations, and/or individuals are able to work together to achieve a common goal. For computer systems, interoperability is typically defined in terms of *syntactic interoperability* and *semantic interoperability*. Syntactic interoperability relies on specified data formats, communication protocols, and the like to ensure communication and data exchange. The systems involved can process the exchanged information, but there is no guarantee that the interpretation is the same. Semantic interoperability, on the other hand, exists when two systems have the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results via deference to a common information exchange reference model. The content of the information exchange requests are unambiguously defined: what is sent is the same as what is understood. More formally, semantic interoperability of data categories C_1 and C_2 is the capability of two annotation consumers to interchange annotation a_1 using C_1 and annotation a_2 using C_2 via a function f that maps C_1 to C_2, such that an analysis of C_2 is identical to the analysis of $f(C_1)$; that is, an analysis should produce the same result for two different but interoperable data categories.

For language resources, the focus is increasingly on semantic rather than syntactic interoperability. That is, the critical factor is seen to be the accurate and consistent interpretation of exchanged data rather than the ability to process it immediately without modification to its physical format. The reasons for this are several, but first and foremost is the existence of large amounts of legacy data in varied syntactic formats, coupled with the continued production of resources representing linguistic information in varied, but mappable, ways. Indeed, to ensure interoperability for language resources, the trend in the field is to specify an *abstract data model* for structuring linguistic data to which syntactic realizations can be mapped, together with a mapping to a set of *linguistic data categories* that communicate the information (linguistic) content. In the context of language resources, then, we can define syntactic interoperability as the ability of different systems to process (read) exchanged data either directly or via trivial conversion. Semantic interoperability for language resources is virtually the same as for software systems: it can be defined as the ability of systems to interpret exchanged linguistic information in meaningful and consistent ways reference to a common set of reference categories.

Semantic interoperability for linguistic annotation has proven to be more

elusive than syntactic interoperability. As early as the 1990s, efforts were devoted to establishing standard sets of data categories, most notably within the European EAGLES/ISLE project[19], which developed standards for morpho-syntax, syntax, sub-categorization, text typologies, and others. However, none of these standards has achieved universal acceptance and use. Recent large-scale efforts addressing standardization of data categories include those within ISO TC37 SC4 (Language Resource Management), which in 2004 proposed a registry accommodating the needs of linguistic annotation (Ide and Romary 2004) and subsequently implemented ISOcat(Kemps-Snijders et al. 2009), an online repository that is accessible and extensible with new data categories by the community. Recently, the ISOcat categories relevant for linguistic annotation were migrated to the *CLARIN Data Concept Registry*[20]. Other efforts include OLiA (Chiarcos 2012), a repository of annotation terminology for various linguistic phenomena intended to apply across multiple languages, and the Web Service Exchange Vocabulary (Ide et al. 2014b) under development within the Language Applications (LAPPS) Grid project (Ide et al. 2014a).

Despite these repeated efforts, there is at the present time no universally accepted set of categories, nor even agreement on what the categories should be. There is, however, some consensus, at least among schemes intentionally tailored to the needs of common NLP tools, which rely on some relatively common practices that have evolved over the years. These commonalities typically refer to attribute types, such as "part-of-speech", "constituent", "semantic role", and "relation" and leave open the range of valid values. This avoids some of the nastier kinds of mapping problems by pushing off problems of harmonization among specific values to another phase or mechanism; for example, tools may be required to provide metadata about the itemized tagsets they input and/or output (e.g., the Penn Treebank part of speech tags or PropBank scheme of semantic role assignment) that can be checked for consistency at run time. Other types of annotation have a fairly consistent (or easily mappable) set of categories, such as nounchunk and verbchunk, coreference (mentions, representative), common subsets of named entities (person, organization, location, date), dependencies (head, dependent), etc. Full consensus on linguistic categories and values is unlikely to be achieved any time soon, if at all; as with syntactic interoperability, the best path may be to find means to allow flexibility while maintaining the ability to map among categories.

[19] http://www.ilc.cnr.it/EAGLES96/browse.html

[20] https://openskos.meertens.knaw.nl/ccr/browser/

6 Conclusion

At this time, there is convergence within the community of means to achieve annotation interoperability and a general willingness to pursue and ensure it. However, it is difficult to identify an obvious solution or even a clear path to follow in order to fully achieve it. New technologies will likely emerge that may affect the way we approach the interoperability problem, much as the development of the Semantic Web and its supporting RDF/OWL format have impacted data models for annotations over the past fifteen years. In the meantime, the plodding progress in pursuit of interoperability that has been made over the past three decades will continue, inching toward a solution that is as yet only distantly visible.

7 Captions

Figure 09.01: Different formats for part-of-speech annotation produced by several tools.

References

Bentivogli, Luisa, Pamela Forner, and Emanuele Pianta. 2004. "Evaluating Cross-language Annotation Transfer in the MultiSemCor Corpus." *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bird, Steven, and Mark Liberman. 2001. "A formal framework for linguistic annotation." *Speech Communication* 33 (1-2): 23–60.

Bond, Francis, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. "Japanese SemCor: A Sense-tagged Corpus of Japanese." *The 6th International Conference of the Global WordNet Association (GWC-2012)*. Matsue.

Bray, Tim, Jean Paoli, C.M. Sperberg-McQueen, Eve Maler, Francois Yergeau, and John Cowan. 2006, September. "Extensible Markup Language (XML) 1.1 (Second Edition)." W3C Recommendation, W3C - World Wide Web Consortium.

Charniak, Eugene. 2000. "A Maximum-entropy-inspired Parser." *Proceedings of the 1st North American Chapter of the Association for Compu-*

tational Linguistics Conference, NAACL 2000. Stroudsburg, PA, USA: Association for Computational Linguistics, 132–139.

Chiarcos, Christian. 2012. "Ontologies of linguistic annotation: Survey and perspectives." *8th International Conference on Language Resources and Evaluation (LREC2012*. 303–310.

Church, Kenneth Ward. 1988. "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text." *Proceedings of the Second Conference on Applied Natural Language Processing*, ANLC '88. Stroudsburg, PA, USA: Association for Computational Linguistics, 136–143.

Clear, Jeremy H. 1993. "The Digital Word." Chapter The British National Corpus of , edited by George P. Landow and Paul Delany, 163–187. Cambridge, MA, USA: MIT Press.

Collins, Michael. 2003. "Head-Driven Statistical Models for Natural Language Parsing." *Comput. Linguist.* 29 (4): 589–637 (December).

DeRose, Steven J. 1988. "Grammatical Category Disambiguation by Statistical Optimization." *Computational Linguistics* 14 (1): 31–39 (January).

Erjaveç, Tomaž, and Nancy Ide. 1998. "The Multext-East Corpus." *Proceedings of First International Conference on Language Resources and Evaluation*. 971–974.

Ferrucci, David, and Adam Lally. 2004. "UIMA: An architectural approach to unstructured information processing in the corporate research environment." *Natural Language Engineering* 10 (3-4): 327–348.

Garside, Roger. 1987. "The CLAWS word-tagging system." In *The Computational Analysis of English*, edited by Roger Garside, Geoffrey Leech, and Geoffrey Sampson, 30–41. London: Longman.

Greene, Barbara B., and Gerald M. Rubin. 1971. *Automatic Grammatical Tagging of English*. Department of Linguistics, Brown University.

Grishman, R. et al. 1998. "The Tipster Annotation Architecture." *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98. Stroudsburg, PA, USA: Association for Computational Linguistics.

Grishman, Ralph, and Beth Sundheim. 1995. "Design of the MUC-6 Evaluation." *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95. Stroudsburg, PA, USA: Association for Computational Linguistics, 1–11.

Ide, Nancy. 1998. "Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora." *Proceedings of the First International Language Resources and Evaluation Conference*. 463–70.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. "XCES: An XML-based Encoding Standard for Linguistic Corpora." *Proceedings of the Second International Language Resources and Evaluation Conference (LREC'00)*.

Ide, Nancy, James Pustejovsky, Christopher Cieri, Eric Nyberg, Di Wang, Keith Suderman, Marc Verhagen, and Jonathan Wright. 2014a, may. "The Language Application Grid." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA).

Ide, Nancy, James Pustejovsky, Keith Suderman, and Marc Verhagen. 2014b. "The Language Application Grid Web Service Exchange Vocabulary." *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT)*. Dublin, Ireland.

Ide, Nancy, and Laurent Romary. 2004. "A Registry of Standard Data Categories for Linguistic Annotation." *Proceedings of the Fourth International Language Resources and Evaluation Conference (LREC'04)*. Lisbon, Portugal, 135–138.

Ide, Nancy, and Keith Suderman. 2014. "The Linguistic Annotation Framework: a standard for annotation interchange and merging." *Language Resources and Evaluation* 48 (3): 395–418.

Ide, Nancy, Keith Suderman, and Brian Simms. 2010, May. "ANC2Go: A Web Application for Customized Corpus Creation." *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC)*. Paris: European Language Resources Association.

Ide, Nancy, and Jean Véronis. 1994. "MULTEXT: multilingual text tools and corpora." *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, Volume I. Kyoto, Japan, 588–592.

ISO. 2012. Language Resource Management - Linguistic Annotation Framework. ISO 24612.

Iso, ISO. 1986. Information processing-Text and office systems – Standard Generalized Markup Language (SGML).

Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2009. "ISOCat: Remodelling Metadata for Lan-

guage Resources." *International Journal of Metadata, Semantics and Ontologies* 4 (November): 261–276.

Kuĉera, Henry, and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Providence, RI, USA: Brown University Press.

Landes, S., C. Leacock, and R.I. Tengi. 1998. "Building semantic concordances." In *WordNet: An Electronic Lexical Database*, edited by C. Fellbaum. Cambridge (Mass.): The MIT Press.

Lupu, Monica, Diana Trandabăţ, and Maria Husarciuc. 2005, July. "A Romanian SemCor Aligned to the English and Italian MultiSemCor." *1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School, Proceedings*. Cluj-Napoca, Romania, 20–27.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. "The Penn Treebank: Annotating Predicate Argument Structure." *Proceedings of the Workshop on Human Language Technology*. Stroudsburg, PA, USA: Association for Computational Linguistics, 114–119.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics* 19 (2): 313–330 (June).

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection." *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association, 1659–1666.

Pustejovsky, James, and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications*. O'Reilly.

Verhagen, M., K. Suderman, D. Wang, N. Ide, C. Shi, J. Wright, and J. Pustejovsky. 2015. "The LAPPS Interchange Format." *Revised Selected Papers of the Second International Workshop on Worldwide Language Service Infrastructure WLSI (2015)*. Kyoto, Japan: Springer International Publishing, 33–47.

Weaver, Warren. 1949; rpt. 1955. "Translation." In *Machine Translation of Languages*, edited by William N. Locke and A. Donald Boothe, 15–23.

Cambridge, MA: MIT Press. Reprinted from a memorandum written by Weaver in 1949.

Zipser, Florian, and Laurent Romary. 2010. "A model oriented approach to the mapping of annotation formats using standards." *Workshop on Language Resource and Language Technology Standards, LREC 2010.* La Valette, Malta.