Community Standards for Linguistically-Annotated Resources

Nancy Ide, Nicoletta Calzolari, Judith Eckle-Kohler, Dafydd Gibbon, Sebastian Hellman, Kiyong Lee, Joachim Nivre, and Laurent Romary

1 Introduction

In some senses, the development of standards for representing linguistically-annotated data in electronic form has been the thorn in the side of language resource creation and use for over thirty years, since the mid-1980s when the use of electronic language data became widespread within the computational linguistics and humanities computing communities. Generally, standardization for representing language resources deals with two phenomena: the *representation format* (syntax) and the *data categories* used to identify linguistic phenomena in the resource. Each poses its own problems for standardization, although standardization of linguistic categories is substantially more problematic because of (sometimes subjective) differences in definitions, granularity, theoretical orientation, etc.

Standardization of both physical formats and linguistic categories has been addressed repeatedly and often over the past thirty years[1], during which steady changes in technology have continuously impacted standardization efforts, by both making dissemination and large-scale community involvement easier and repeatedly supplanting (while improving upon) implementation options. Recent times have seen considerable convergence of practice for representation format as well as more general agreement on the means and mechanisms by which to provide "semantic interoperability" [47] via standardized data categories among resources. Nevertheless, to date, no single, universally accepted set of best practice guidelines for either of these concerns for the creation of linguistically-annotated resources exists.

There has traditionally been a division of opinion about the need for such standards: on the one hand, the need to enable reusability and sustainability of language resources via standards was evident to many, while others felt that standards were

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

[1] Note that until roughly 2001, the separation of physical format and linguistic information was typically not taken into account in the development of standards for language resources.

unnecessary and/or inhibiting, or would arise *de facto* from ongoing work. At the extreme, these attitudes are manifested in two opposing approaches to standards development: a top-down approach, which seeks to define a standard more or less *a priori*, possibly anticipating needs even before they arise in practice; and a bottom-up approach driven by the needs of specific projects and software. Most focused standards development efforts fall somewhere in between these two extremes but closer to the top-down approach, as opposed to *de facto* standards that are project-driven and eventually adopted "because they are there".

This chapter provides a broad overview of the state-of-the-art in standards development for language resources, beginning with a brief historical overview to serve as context. It describes in some detail several current, major efforts that define the standardization landscape for language resources today, with the aim of outlining their differences and commonalities and, more generally, identifying the progress that has been made to date as well as the obstacles to definitive standardization. In addition to describing standards that are most applicable to linguistic annotation of text, we include a section that overviews considerations and alternatives for spoken data. We also overview a widely-used and influential *de facto* standard and consider its role in standards development. Finally, we provide an assessment of the standards landscape and the options available to current and future creators of linguistically-annotated resources.

2 History

The need for standards for representing language resources has been acknowledged since the 1980s, when schemes for representing textual material in electronic form began to proliferate. Most schemes were developed by specific groups or individuals for a specific purpose, and as a result they were typically idiosyncratic and incompatible for use by other projects or with different software than that for which they were originally designed. The situation was exacerbated by the practices of software vendors and electronic publishers, who often developed proprietary formats as part of a business strategy to benefit a particular company. At the same time, as the use of electronic language data became increasingly widespread within the computational linguistics and humanities computing communities, the drawbacks of language data that could not be reused and was not sustainable were increasingly evident. Thousands of hours were spent converting data represented in one format to another that would work for a different purpose or with different software, or, worse, recreating the same resources to suit a particular need. Thus "reusability" for language data became a mantra in the late 80s and early 90s, especially in Europe; in recent years, the term "interoperability", which applies broadly to both data and software, has become the primary watchword.

Any history of international standardization efforts for encoding texts in electronic form must begin with the Text Encoding Initiative (TEI), which was formally established in 1987 at a meeting held at Vassar College in Poughkeepsie, New York, funded by the US National Endowment for the Humanities (NEH). The meeting was attended by thirty-five representatives of major projects and organizations from around the world, all of whom contributed to devising the "Poughkeepsie Principles"[2], a summary document that outlined the basic design goals and working principles for the encoding guidelines to be created by the TEI. The primary goal of the effort was stated to "provide explicit guidelines that define a text format suitable for data interchange and data analysis; the format should be hardware and software independent, rigorous in its definition of textual objects, easy to use, and compatible with existing standards." To this end, the attendees agreed that the TEI's encoding guidelines would consist primarily of a set of tags represented using the syntax of the recently introduced Standard Generalized Markup Language (SGML)[3]–itself a bold move at the time–accompanied by a description of their meanings and interrelationships.[4]

The TEI's focus was on "machine-readable texts intended for literary, linguistic, historical, or other textual research", and as such, the initiative's activity centered, and largely continues to center, on the needs of humanities-based research. The first official edition of the TEI Guidelines, which appeared in 1994 [111], included means for detailed encoding of phenomena in historical manuscripts, verse, drama, print dictionaries, and terminological databases, together with extensive mechanisms for linkage and alignment, indication of certainty and responsibility, transcription of primary sources, critical apparatus, and the like. Additional tagsets were defined for less specifically humanities-oriented phenomena such as graphs, trees, networks and feature structures, but because of the focus on humanities text types, the TEI Guidelines were and continue to be used primarily by humanities scholars. However, the impact of the TEI Guidelines as a pioneering effort can be seen to this day throughout the text and data encoding world. See Section 3 for an overview of the current TEI Guidelines and future development plans.

The philosophy underlying development of the TEI Guidelines was to accommodate a wide variety of potential needs, and most of the specifications were developed prior to their application in real data, making it a fundamentally top-down exercise. This meant that the TEI Guidelines provided multiple alternative ways to encode the same phenomenon, which to some extent undermined the original goal of standardization. This, together with the focus on humanities data, motivated creation of the Corpus Encoding Standard [24], an application of the TEI Guidelines developed in 1994 for representing linguistically annotated corpora. The CES limited the range of options for encoding the same phenomenon in order to identify a single, standard representation, and extended the TEI mechanisms for more comprehensive coverage of phenomena such as part of speech and syntax, parallel text alignment, and transcription of spoken data. The CES also defined and recommended the use of

[2] The Poughkeepsie Principles together with an accounting of the founding assumptions and sponsors of the TEI are available at http://www.tei-c.org/Vault/ED/edp01.htm#b2b1b3b3b3

[3] SGML was formally adopted as an ISO standard in 1986; see [59]

[4] The TEI Guidelines were later converted to the Extensible Markup Language (XML) which superseded SGML in the mid-1990s and whose design was influenced by work undertaken in the TEI project.

standoff markup[5] (see Part I -Ch. III - Sect. 3.2.4), which was subsequently adopted in the DARPA TIPSTER Architecture [41] and the General Architecture for Text Engineering (GATE) framework [27], and is now widely accepted as best practice for linguistically annotated resources.

EAGLES (Expert Advisory Group for Language Engineering Standards) was established as an EU-funded project in 1993 to provide standards, common guidelines, and best practice recommendations for large-scale language resources (e.g., text corpora, computational lexicons and speech and multimodal resources), together with means for manipulating and evaluating these resources via computational linguistic formalisms, markup languages, and software.[6] The effort was extensive and published a wide variety of standards, including standard corpus and text typologies, standards for encoding spoken data, and standards for linguistic software development.[7] Two of the most widely-used and influential EAGLES standards are the CES (described above) and the extensive EAGLES guidelines for morpho-syntactic annotation of corpora and lexicons. The latter define a common core of morpho-syntactic distinctions applicable to all Western and Eastern European languages, together with a layered set of additional, optional language-specific distinctions. In contrast to the top-down approach of the TEI, both the CES and the EAGLES morpho-syntactic specifications were developed in the course of their application to large-scale corpora and lexicons in the EU-funded MULTEXT [57] and MULTEXT-EAST[8] projects. The existence of resources embodying these standards led to widespread adoption and enabled their influence on later standards development, including the morpho-syntactic data categories in ISO-12620 (ISOcat–see Section 4.3) and the ISO-26212 Linguistic Annotation Framework (LAF) ([56]; see also Section 4.1), which drew from the CES and its XML instantiation, XCES [46].

Throughout the 1980s and 90s, standards for linguistic annotation specified both a prescribed physical format and fixed set of content categories or “linguistic labels”.[9] So, for example, standards such as the TEI Guidelines and the XCES used XML as the physical format for annotations, but also standardized the *labels* used to describe linguistic objects in XML element names and XML attribute names and values. However, in 2001, two separate efforts introduced standards that abstract away from file formats, coding schemes, and user interfaces in order to provide a logical basis for linguistic annotations and thus allow for flexibility in the physical rendering of annotated data. Annotation Graphs (AG) [7] provided a formal framework for representing linguistic annotations of time series (spoken) data by specifying means to define a set of graphs, each representing an individual annotation layer, whose nodes are anchored at time stamps and labeled edges that identify spans of data and provide their linguistic labels. Similarly, ISO LAF introduced a

[5] Originally called “remote markup”–see http://www.cs.vassar.edu/CES/CES1-5.html.

[6] ISLE (International Standards for Language Engineering), a standards-oriented transatlantic initiative, was established in 2000 as a continuation of EAGLES.

[7] EAGLES Guidelines are still available at http://www.ilc.cnr.it/EAGLES/browse.html

[8] http://nl.ijs.si/ME/

[9] Part I - Ch. III - Sect. 2 in this volume provides a history of the development of standards for physical format.

graph-based model for defining one or more inter-connected layers of linguistic annotations over data in any medium [49] (see Section 4.1). The notable departures in both of these standards were (1) the definition of an *abstract annotation model* that could be serialized in any of a variety of physical formats; and (2) separation of the specification of annotation content categories from specification of the physical format. The AG framework left the choice of content categories to the annotator, while ISO LAF provided means to link nodes in the graph of annotations to content categories defined in one or more web-based repositories (initially, ISOcat). As such, ISO LAF was a pre-cursor of the RDF/OWL "Linguistic Linked Data" model that is currently gaining attention as a means to inter-link linguistically-annotated resources in the Semantic Web.[10]

After 2001, the graph-based model substantially influenced the development of new linguistic annotation schemes and helped achieve greater syntactic interoperability among annotated resources (i.e., ability of different applications to handle different physical formats, often via trivial mapping). At the same time, semantic interoperability, which would enable systems to understand the meaning of annotation labels and features from other sources, remains an elusive goal. Following the model of ISOcat, current efforts focus on the development of repositories of linguistic terms to serve as a reference point for linguistic annotations, so that terminology is unambiguously and consistently defined and common concepts are identified via mapping to terms in the repositories. Other efforts include the General Ontology of Linguistic Description (GOLD) [30], the LAPPS Web Service Exchange Vocabulary [48], and the OLiA ontologies [25], which formalize numerous annotation schemes for morphosyntax, syntax and higher levels of linguistic description, and provide a linking to the morphosyntactic profile of ISOcat, GOLD, and other terminology repositories. In addition, an RDF/OWL-based standard, the NLP Interchange Format (NIF), has recently been developed to achieve greater interoperability among tools, language resources and annotations via Linked Data technologies and practices. NIF addresses both syntactic and semantic interoperability, relying on a *NIF Core Ontology* to define structural concepts and a selection of ontologies for referencing common NLP terms and concepts (see Section 7).

3 Broad-based Standards: The Text Encoding Initiative

As noted in the previous section, the TEI Guidelines for encoding textual data were first published in 1994 and have undergone two major updates since, the last one published in 2003. The Guidelines are still widely used for encoding humanities texts and have influenced, both directly and indirectly, the development of representation standards for language resources since its introduction.

The current TEI Guidelines address a wide range of textual genres, including manuscripts, drama, speech transcriptions, dictionaries, and others. A TEI docu-

[10] See Part I, Chap. 3, Sect. 5.2.

ments top-level structure, depicted in Figure 1, combines a mandatory “header” providing extensive metadata for the document with the actual content. The content (“text”) can be further divided into “front”, “body” and “back” sections, which allow for encoding the source document together with additional resources such as a table of contents, bibliographies, or a timeline (for example, in a speech transcription). The TEI header is perhaps one of the most influential of the initiative’s developments, as it provided the first standard means to comprehensively identify the provenance, creation practices, attribution information, etc. for machine-readable documents.

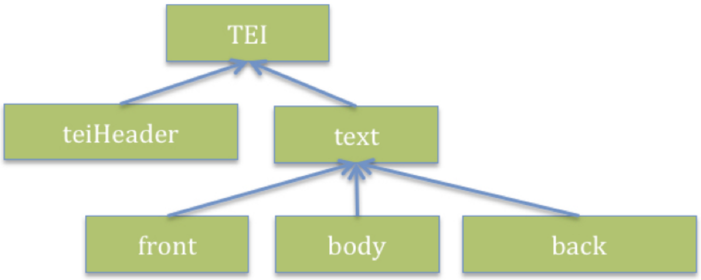


Fig. 1 TEI document architecture

The TEI vocabulary provides several different means to encode a document, which fall into the following main categories:

- Description of the structure of a text by means of the generic `<div>` element, which can be used recursively to describe a hierarchy of textual divisions
- Organization of the content into paragraph-level objects such as paragraphs, lists, figures, tables, etc.
- Inline annotation elements to mark up specific linguistic segments (such as highlighted objects and foreign expressions) or reference to entities (such as names, dates, and numbers)
- Domain-specific constructs for dealing with turns in speech transcription, dictionary entries, etc.
- General-purpose representation objects such as bibliographical descriptions

All these elements are part of the TEIs reference framework, from which a given project can make a selection of applicable components depending on its needs and objectives. The TEI Guidelines provide mechanisms to express such customizations, as described in the next section.

3.1 The TEI specification framework

The TEI Guidelines can be regarded from two different perspectives. First, as the basis of an XML representation format, they provide the technical constraints to control the validity of TEI-conformant document instances. Second, they are delivered with an extensive prose description that informs users about the logic of the guidelines as well as the most appropriate way(s) to use them to represent specific textual phenomena.[11] These two views of the TEI Guidelines, rather than comprising two separate components, are integrated in a single specification from which each view can be automatically generated. This mechanism, following the tenets of "literate programming" [72], is based in an underlying specification language named ODD (One Document Does it all), which is itself expressed in TEI.

In the TEI infrastructure, each element is defined as an ODD specification providing all the necessary information both to control its (XML) syntactic behavior and generate the corresponding documentation including a gloss, a definition, a technical description of the content model, the various attributes it can bear, and one or more examples of usage.

The TEI framework also provides two mechanisms central to ensuring the Guidelines' global coherence: *classes* and *modules*. Two types of classes are defined: *attribute classes*, which group together attributes used in the same way across various elements,[12] and *model classes*, which group together elements that have related semantics and occur at the same locations in a document. The latter provide means to simplify the expression of content models and facilitate the customization process for adding or removing an element from a class. Modules are more global objects, intended to group together coherent sets of elements designed for a similar purpose. For example, all the elements specific to dictionary encoding are grouped together in a single module.

3.2 TEI and linguistic annotations

In the remainder of this section, we show several TEI constructs and mechanisms for representing phenomena in linguistically-annotated corpora. Where applicable, we refer to ongoing ISO/TC 37 SC4 activities to illustrate how a possible transition to more elaborate annotation schemas, or mappings from basic TEI representations to other annotation schemas, could be implemented.

The TEI Guidelines provide mechanisms for both inline and stand-off annotations (see Chapter III, section 3, for a discussion of standoff vs. inline annotation). Inline annotation has traditionally been the primary TEI mechanism for identifying

[11] In particular, the TEI guidelines contain a wealth of examples for each element and the major constructs they allow.

[12] For instance, the class *att.global*, which contains general purpose attributes such as the W3Cs @XML:ID and @XML:LANG and the TEIs generic @N (for local numbering) and @REND (for rendering information).

entities within a text. The TEI vocabulary contains a wealth of elements for inline annotation of, for example, numbers, measurements, dimensions, temporal expressions, and geographical coordinates. All elements may be associated with attributes to normalize tagged content according to established standards (e.g., ISO 3166 for country codes). The TEI also includes a variety of elements to tag names and other referring expressions, either at a generic level or specifically for person or place names and components thereof. Finally, there are several combinations of elements for tagging structured portions of a text, such as bibliographical references, formulas, tables, or graphics.

As opposed to the comprehensive XML vocabulary provided for inline annotation of documents, stand-off annotation in the TEI guideline relies upon a number of generic constructs that can be easily applied to various annotation scenarios. A generic `<span>` element enables reifying any type of segment in order to supply further annotations. The `<span>` element is conceptually close to the notion of "markable"(see [22]) in many annotation schemes and is also parallel to the REGION component in ISO 24612 (LAF). It may also be used to reify more abstract components in an annotation scheme, as described below in the case of ISO 24611 (MAF). A `<link>` element allows the encoder to express a relation between any two objects within a document or across various documents. For instance, it can be used to represent multilingual alignments [105] and complex syntactic annotations [40].

Beyond these generic mechanisms, the TEI Guidelines provide several technical components intended to facilitate the precise annotation of linguistic content:

- A set of general pointing attributes grouped together within an attribute class (att.global.linking) that is used with numerous elements to express similarity (@CORRESP, @SAMEAS), difference (@EXCLUDE, @SELECT) or temporal synchronisation (@SYNCH)
- A comprehensive module, also published as an ISO standard[13], to describe feature structures, constraints on feature structures and libraries of features and feature structures
- The native integration of data category attributes, allowing one specific annotation to align with a data category (@DATCAT) or a value (@VALUEDATCAT) in the ISO data category registry

3.3 Relation to ISO standards

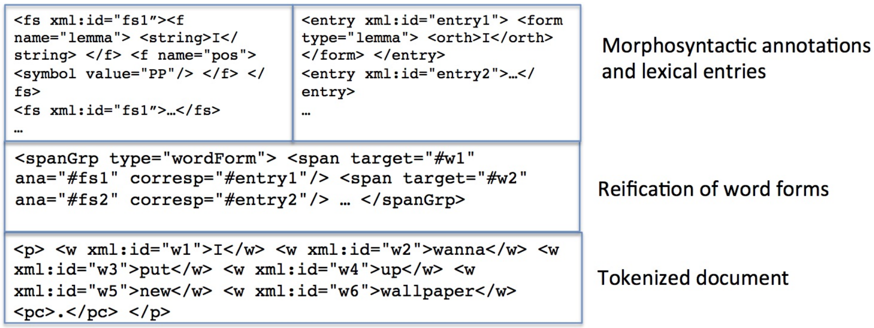
One of the early proposals of ISO/TC 37/SC 4 was to outline a possible standard for morphosyntactic annotation (also referred to as part-of-speech annotation). Morphosyntactic annotation is typically the first level of linguistic abstraction level over a corpus, and, depending on the language of the primary data, the tool used to anno-

[13] ISO 24610-1:2006 Language resource management – Feature structures – Part 1: Feature structure representation

tate, and the theoretical underpinnings of the annotation scheme, it can vary enormously in structure and complexity. To deal with the complex issues of ambiguity and determinism in morphosyntactic annotation, ISO 24611 makes a distinction between *tokens*, which represent a surface segmentation of the source, and *word forms* that represent lexical abstractions associated to groups of tokens. Each of these can be represented as a simple sequence or a local graphs (e.g. multiple segmentations, ambiguous compounds, etc.), and any *n*-to-*n* combination can hold between word forms and tokens–i.e., one token may correspond to several word forms, and vice versa. ISO 24611 provides a standard TEI-based serialization that implements the various components of the MAF meta-model, as illustrated in Figure 2. Specifically:

- The token level is implemented by means of both <w> for lexical tokens and <pc> for punctuation. For ease of reference, every instance of both elements is required to be uniquely identified by means of the @XML:ID attribute;
- The <span> element serializes the wordForm component and by means of the @ANA and @CORRESP attributes, refer to the morphosyntactic annotation and the associated lexical entry, respectively;
- Morphosyntactic annotations are represented as a feature structure encoded according to the ISO-TEI feature structure standard;
- The reference lexical entry associated with a wordForm uses the TEI <entry> element [58], with a further compliance constraint to ISO 24613 (see [74]).

Both the MAF model and the model of ISO standard 24615 (SynAF) for syntactic annotations[14] can be implemented using the basic mechanisms of the TEI. More generally, the specification platform of the TEI Guidelines make it easy to incorporate an external vocabulary within a TEI-based customization.[15]



```
<fs xml:id="fs1”><f
name="lemma"> <string>I</
string> </f> <f name="pos">
<symbol value="PP"/> </f> </
fs>
<fs xml:id="fs1”>…</fs>
…
```

```
<entry xml:id="entry1"> <form
type="lemma"> <orth>I</orth>
</form> </entry>
<entry xml:id="entry2">…</
entry>
…
```

```
<spanGrp type="wordForm"> <span target="#w1"
ana="#fs1" corresp="#entry1"/> <span target="#w2"
ana="#fs2" corresp="#entry2"/> … </spanGrp>
```

```
<p> <w xml:id="w1">I</w> <w xml:id="w2">wanna</w> <w
xml:id="w3">put</w> <w xml:id="w4">up</w> <w
xml:id="w5">new</w> <w xml:id="w6">wallpaper</w>
<pc>.</pc> </p>
```

Morphosyntactic annotations and lexical entries
Reification of word forms
Tokenized document

Fig. 2 ISO 24611 serialization of MAF

[14] See the implementation in the Polish National corpus [98]

[15] See for instance [104] for introducing TBX entries within a TEI document.

3.4 Linguistic annotation projects based upon the TEI Guidelines

The TEI Guidelines have been used by numerous projects to represent linguistic annotations. The following outline a few representative cases.

Most of the morpho-syntactically annotated corpora using the TEI Guidelines are not compliant to the two-level annotation model of ISO 24611, instead merging the token and word-form levels by directly attaching lemma and part-of-speech annotations to the `<w>` element in a tokenized text with the @LEMMA (or @LEMMAREF) attributes, as shown in the MorphAdorner[16] output in Figure 3.

```
<l>
 <w lemma="allow" ana="#vvb" reg="Allow"
     xml:id="A01055-004840">Allow</w>
 <w lemma="thy" ana="#po21" reg="thy"
      xml:id="A01055-004850">thy</w>
 <w lemma="scene" ana="#n2" reg="Scenes"
      xml:id="A01055-004860">Sceanes</w>
 <w lemma="and" ana="#cc" reg="and"
      xml:id="A01055-004870">and</w>
 <w lemma="stile" ana="#n1" reg="Style"
      xml:id="A01055-004880">Stile</w>
 <pc xml:id="A01055-004890">:</pc>
 <w lemma="ay" ana="#uh" reg="ay"
      xml:id="A01055-004900">I</w>
 <pc xml:id="A01055-004910">,</pc>
 <w lemma="as" ana="#c-acp" reg="as"
      xml:id="A01055-004920">as</w>
 <w lemma="a" ana="#dt" reg="a"
      xml:id="A01055-004930">a</w>
 <w lemma="friend" ana="#n1" reg="friend"
      xml:id="A01055-004940">friend</w>
</l>
```

Fig. 3 MorphAdorner output

MULTEXT-East[17] has, since its beginnings in 1994, used the TEI Guidelines as a reference for the encoding of its textual content, providing a strategy adopted in several subsequent projects.[18] The following example illustrates the encoding principles behind the JOS corpus [116], with a sentence tokenized and part-of-speech-tagged by means of the `<w>` element, together with dependency annotations encoded using various `<link>` elements, as shown in Figure 4. This is an example of the "hybrid standoff" annotation strategy described in Chapter III (section 3.2.4).

[16] See http://morphadorner.northwestern.edu, with the annotation tagset described in http://panini.northwestern.edu/mmueller/nupos.pdf.

[17] http://nl.ijs.si/ME/

[18] See http://nl.ijs.si/jos/, http://eng.slovenscina.eu/, and http://nl.ijs.si/imp/

```
<s xml:id="F0020003.557.2">
 <w xml:id="F0020003.557.2.1" lemma="ta" msd="Zk-sei">To</w><S/>
 <w xml:id="F0020003.557.2.2" lemma="biti" msd="Gp-ste-n">je</w>
 <term type="sloWNet" sortKey="kraj" subtype="missing_hyponym"
           key="ENG20-08114200-n">
   <w xml:id="F0020003.557.2.3" lemma="turisticen"
           msd="Ppnmein">turisticen</w>
   <w xml:id="F0020003.557.2.4" lemma="kraj" msd="Somei">kraj</w>
  </term>
   <c xml:id="F0020003.557.2.5">.</c><S/>
</s>
<linkGrp type="syntax" targFunc="head argument"
           corresp="#F0020003.557.2">
 <link type="ena" targets="#F0020003.557.2.2 #F0020003.557.2.1"/>
 <link type="modra" targets="#F0020003.557.2 #F0020003.557.2.2"/>
 <link type="dol" targets="#F0020003.557.2.4 #F0020003.557.2.3"/>
 <link type="dol" targets="#F0020003.557.2.2 #F0020003.557.2.4"/>
 <link type="modra" targets="#F0020003.557.2 #F0020003.557.2.5"/>
</linkGrp>
```

Fig. 4 Example sentence from jos100k: “To je turisticen kraj.” , lit. “It is a tourist place.”

Software support for linguistic annotation with TEI includes the TXM annotation tool[19], which uses TEI mechanisms and an extension for encoding linguistic annotations in its pivot source format (Heiden, 2010). The TXM customization is based upon basic segmentation of texts into sentences (<s>) and tokens (<w>) together with <interp> for all additional annotations. The TXM import environment also implements a TEI-TXM standoff schema using the TEI <linkGroup> and <link> elements in standalone TEI text files, which point back to <w> elements in TEI-encoded texts.[20]

Finally, it is important to mention the recent trend in several projects based upon spoken data of adopting the TEI Guidelines as a dissemination format independent of the formats used by various tools available for the transcription and annotation of spoken data (cf. handbook). Such annotation usually comprises basic interlocution annotation (cf. Corpus français) up to complex dialogue-act representation (in line with the recently published ISO standard, see [16]). This has led to a new ISO project (24624) to standardize the representation of speech transcription, based on the corresponding TEI chapter and the customization work described in [110].

[19] See also Chapter III, section 3.2.4 for a description of the MMAX2 annotation tool.

[20] The reference specification of the TEI-based TXM pivot format is available at http://txm.sourceforge.net/wiki/index.php/XML-TXM.

3.5 Summary

Over the years, the TEI Guidelines have become the reference standard for encoding primary sources in the humanities. As we have seen, the guidelines provide various means to enrich documents with linguistic annotations, and humanities projects are typically content to remain within the TEI framework when their annotations are strongly related to the nature of source (oral, epigraphic, manuscript, etc.). At the same time, the TEI's inline and stand-off annotation mechanisms can be mapped to existing or developing international standards, either natively or by using the TEI customization mechanisms. Given the stabilization of encoding practices within the TEI user community (e.g., the use of hybrid standoff based on TEI-encoded tokens) together with the possibility to adapt and test external annotation schemes within the TEI architecture, there is strong potential achieve convergence between the TEI Guidelines and international efforts such as those carried out within ISO/TC 37 SC4 (cf. [74]).

4 Ongoing Efforts: ISO Standards for Language Resource Management

In 2002, a sub-committee of technical committee (TC) 37, *Terminology and other Language and Content Resources* was formed within the International Standards Organization (ISO) to propose, draft, review, and revise documents describing standard practices for Language Resource Management (LRM) to be eventually published by ISO as international standards. In the twelve years since its formation, numerous scholars and researchers have been involved in developing specifications for these standards, which include an overall architecture for representing annotated corpora and representations for several different types of linguistic annotation. To date, the sub-committee (formally known as ISO TC37 SC4) has published twelve international standards covering different aspects of LRM, and several major efforts are ongoing. Overviews of ISO TC37 SC4 standards and activities appear in [54, 78].

Within SC4, six working groups have been so far established, which have so far produced fifteen international standards:

- ISO 24610-1:2006, Language resource management - Feature structures - Part 1: Feature structure representation (FSR)
- ISO 24610-2:2011, Language resource management - Feature structures - Part 2: Feature system declaration (FSD)
- ISO 24619:2011, Language resource management - Persistent identification and sustainable access (PISA)
- ISO 24612:2012, Language resource management - Linguistic annotation framework (LAF)
- ISO 24615:2010, Language resource management - Syntactic annotation framework (SynAF)

- ISO 24614-1:2010, Language resource management - Word segmentation of written texts - Part 1: Basic concepts and general principles (WordSeg-1)
- ISO 24614-2:2011, Language resource management - Word segmentation of written texts - Part 2: Word segmentation for Chinese, Japanese and Korean (WordSeg-2)
- ISO 24611:2012, Language resource management - Morpho-syntactic annotation framework (MAF)
- ISO 24617-1:2012, Language resource management - Semantic annotation framework (SemAF) - Part 1: Time and events (SemAF-Time, ISO-TimeML)
- ISO 24617-2:2012, Language resource management - Semantic annotation framework (SemAF) - Part 2: Dialogue acts (SemAF-DA)
- ISO 24617-4:2014 Language resource management - Semantic annotation framework - Part 4: Semantic roles (SemAF-SR)
- ISO 24617-7:2014 Language resource management - Part 7: Spatial information (ISOspace)
- ISO 24616:2012, Language resource management - Multilingual information framework (MLIF)
- ISO 24613:2008, Language resource management - Lexical markup framework (LMF)
- ISO 24615-1:2013, Language resource management - Syntactic annotation framework (SynAF) - Part 1: Syntactic model[21]

Several additional topics are currently under active development, including discourse structures (ISO DIS 24617-5) and Basic Principles for a Semantic Annotation Framework (ISO DIS 24617-6).

The following sections describe some of the most well-known of the SC4 standards for linguistic annotation.

4.1 Linguistic Annotation Framework

The development of the Linguistic Annotation Framework (LAF) was the first work item established by the sub-committee in order to provide a broad framework for more specific standards for representing linguistic annotations that have been and continue to be developed in other SC4 working groups. The earliest work on LAF involved identifying the fundamental properties and principles for representing linguistic annotations that satisfied the criteria for expressive adequacy, media independence, flexibility, processability, and–perhaps most critically–mappability to the objects and relations in a variety of formats suitable for different tools and applications.

The original design of LAF was outlined in 2001 and later summarized in [50, 51, 54]. It was based on two fundamental principles: (1) adoption of an abstract

[21] This is a version based on ISO 24615:2010 SynAF, with the title changed.

data model that clearly separates annotation structure (the physical format of annotations) and annotation content (the categories or labels used to describe linguistic phenomena; and (2) adoption of *standoff annotation*, in order to preserve the original form and content of the primary data and allow for multi-layered annotations and multiple annotations of the same type. The abstract data model was defined to be an *acyclic di-graph* decorated with *feature structures*; the complete LAF data model includes :

1. a structure for describing media, consisting of *anchors* that reference locations in primary data, and regions defined in terms of these anchors;
2. a *graph structure*, consisting of nodes, edges, and links to regions; and
3. an *annotation structure* for representing linguistic information with feature structures.
4. provision for *URI-based references to linguistic categories* defined in existing repositories as a means to achieve semantic interoperability.

In 2007, the Graph Annotation Format (GrAF) (Ide and Suderman, 2007) was introduced as the XML serialization of the LAF abstract data model; it was subsequently modified slightly in response to input from experience with full-scale implementation in two multi-layered corpora (Open American National Corpus and MASC[22] [45]) and implementations for multi-media data, as well as issues that have arisen in the course of developing the ISO standards for specific annotation types. The ISO standard describing LAF and GrAF is published as ISO 24612:2012[64] (see also [56]).

GrAF is intended to serve as a *pivot format* into and out of which representations of annotations in other formats can be mapped to facilitate interoperability, and not as a stand-alone format on its own. The LAF abstract model that GrAF serializes therefore was used as the basis for development of all other SC4 annotation standards, as well as a standard for encoding lexicons (Lexical Markup Framework (LMF) [35]). See Part I, Chap. III, Sect. 5 for additional description of LAF/GrAF.

Because LAF uses a graph-based model, it is very similar–and in most cases, isomorphic–to many recently-defined formats (see Part I, Chap. III for several examples), including the Linked Data format RDF/OWL. As such, in terms of physical format LAF/GrAF is trivially mappable to most other formats, which represents a major step toward syntactic interoperability. The most important contribution of the standard is likely its fostering of a principled data model–in particular, the graph–as a basis for linguistic annotation schemes.

4.2 ISO SemAF: Semantic Annotation Schemes

ISO has published five semantic annotation schemes as international standards under ISO/TC 37/SC 4/WG 2 Semantic Annotation: SemAF-Time (ISO-TimeML)

[22] See Part II, Chapter I.c.

[62], SemAF-DA [63], SemAF-SR [65], and ISOspace [66]. Each provides an annotation scheme for the markup of specific semantic phenomena: SemAF-Time treats time and event-involving temporal information, while ISO-TimeML is an XML-serialization of SemAF-Time. SemAF-DA treats dialogue acts in everyday language, SemAF-SR the semantic roles of participants in each eventuality, and ISOspace location or motion-related spatial information in text.

Each annotation scheme has two levels: the level of abstract syntax and that of a concrete syntax, which is based on an abstract syntax. The abstract syntax specifies in abstract formal terms how a language, either written or spoken, and sometimes with iamges, is annotated for some particular types of information, whereas a concrete syntax shows how each annotation is represented in an accepted markup language such as XML. Note that an abstract syntax may allow a variety of concrete syntaxes that all represent annotations equivalently. An XML-serialization such as TimeML [99] or SpatialML [87] is an example of a concrete syntax.

4.2.1 ISO-TimeML: Annotation of Time and Events

One of the first and most widely-used ISO standards for language resource annotation is ISO-TIMEML, which provides a set of annotation guidelines for temporal and event-related information. ISO-TimeML [62], introduced in [101], grew out of TimeML [102] and [99]. Both schemes provide an XML-serialization of an annotation scheme for annotating time and event-related information in language. The abstract syntax of ISO-TimeML consists of (1) three types of temporal expressions, all tagged as `<TIMEX3>`, date, duration, frequency, (2) four different types of temporal link, subordinate link, aspectual link, and measure link, tagged as `<TLINK>`, `<SLINK>`, `<ALINK>`, and `<MLINK>`, respectively, and (3) temporal signals, tagged as `<SIGNAL>`.

There are three basic differences between TimeML and ISO-TimeML. First, following LAF [52] and [61], ISO-TimeML adopts standoff annotation instead of inline annotation. Second, TimeML treats event instance, tagged `<EVENT-INSTANCE>`, as a basic entity, but there are no such event instances in ISO-TimeML, for each event or eventuality in ISO-TimeML is understood to be an event instance. Second, temporal durations are often interpreted in ISO-TimeML correctly as referring to time amounts, as in *John taught [three hours]$_t$1 last week*. In ISO-TimeML, such time amounts are linked to events by the measure link `<MLINK>` instead of the temporal link `<TLINK>`.

Most of the attribute names and their possible values in TimeML are adopted by ISO-TimeML. There are, however, two new attributes `@target` and `@pred`: the first one refers to a markable in text and the second one represents the content of a markable. Both TimeML and ISO-TimeML follow ISO 8601 [60] in representing dates and times including durations and time amounts such as `value="P2D"` for *two days*, although there was a strong argument against such an adoption.

Here is an example of annotating the amount of time in ISO-TimeML:

(1) *John traveled for two weeks last December.*

```
<EVENT xml:id="e1" target="#token2" pred="TRAVEL" tense="PAST"/>
<SIGNAL xml:id=′s1" target="#token3" pred="FOR"/>
<TIMEX3 xml:id="t1" target="#token4, #token5" pred="TWO_WEEK"
type="DURATION" value="P2W"/>
<TIMEX3 xml:id="t2" target="#token6, #token7" pred="LAST_DECEMBER"
type="DATE" value="2014-12-XX"/>
<MLINK eventID="#e1" relatedToTime="#t1" relType="MEASURE"/>
<TLINK timeID="#t1" relatedToTime="#t2" relType="DURING"/>
```

Here is an example of annotating time interval in ISO-TimeML:[23]

(2) *We drove to Niagara Falls [*$_{t21}$**three days**$_{t22}$*]*$_{t2}$ *before* **Christmas Day**$_{t3}$.

```
<TIMEX3 xml:id="t2" type="DURATION" value="P3D"
beginPoint="#t21" endPoint="#t22"/>
<TIMEX3 xml:id="t3" type="DATE" value="XXXX-12-25"/>
<TIMEX3 xml:id="t21" target="" type="DATE" value="XXXX-12-22"
temporalFunction="TRcitetUE" anchorTimeID="#t3"/>[24]
```

As the first part of ISO's international standard on semantic annotation, ISO-TimeML has taken up two very important tasks. One task concerns the introduction of the notion of *abstract syntax* vs *concrete syntax* into the specification of an annotation scheme, as motivated by [15]. Another task relates to the construction of a semantics for a semantic annotation scheme. [97] developed an interval-based formal semantics for TimeML and then a slight revised version for ISO-TimeML. Besides this interval-based semantics, ISO-TimeML also contains an event-based formal semantics, which was developed by [14]. Besides these works, there are other efforts to develop formal semantics for TimeML or ISO-TimeML such as [68], [75], and [76].

The current version of ISO-TIMEML (2012) requires further refinement. For example, an expression such as *2 days* does not denote an interval, but rather the length of a temporal interval–the temporal equivalent of a spatial distance (e.g., *2 miles*). Accordingly, the metamodel introduces **amounts of time** as an element distinct from temporal **instances** or **intervals**. ISO-TIMEML then introduces a "measure link". [100] claim that the problem of linking events to amounts of time is resolved simply by introducing a link with the inherent relation type MEASURE that "reifies the role that certain expressions in the language play in measuring over a time". A time-amount expression such as *three hours* can then be subject to the interpretation of a time amount [20].

4.2.2 ISOspace for Spatial Information

ISOspace refers to *ISO 24617-7:2014 Language resource management - Semantic annotation framework - Part 7: Spatial information (ISOspace)*. Its scope goes be-

[23] Copied from [77]

[24] `<TIMEX3 xml:id="t21"/>` may be treated as an element, called *non-consuming tag*, which has no associated markable expression in text, thus the value of its attribute `@target` being empty `""`. See ISOspace [66], A.3.4 Special Section: Non-consuming tags.

yond MITRE's *SpatialML* [79], the previous state-of-the-art standard upon which ISOspace expands, in two respects: first, ISOspace treats motion-involving dynamic spatial information beyond qualitative spatial information, and second, ISOspace provides an abstract syntax on which a variety of concrete syntaxes such as an XML-serialization can be developed to represent annotations.

The abstract syntax of ISOspace consists of a set M of markable expressions, a set of basic entities, a set R of binary links over basic entities, and a set @ of attribute-value assignments to each entity in E and each link in R. Specifically, markable expressions are words, sequences of words or even morphemes which carry information as delimited by the set of basic entities. The set E of basic entities include:

1. spatial entity (se): location: place (pl), *Boston*$_{pl1}$, and path (pa), *[I 90]*$_{pa1}$
2. event (e): motion (m), *drive*$_{m1}$, and non-motion event (e), *lives*$_{e1}$
3. signal (s): spatial signal (ss), *at*$_{ss}$ *home*, motion signal (ms), *from*$_{ms}$ *Seoul*, and measure signal (mes), *[about 8 miles]*$_{mes1}$

The set R of links include: (1) qualitative spatial link (qsLink) (2) orientation link (oLink), (3) move link (moveLink), and (4) measure link (mLink).

Each possible attribute-value assignment in @ is then specified in the form of an XML DTD or a table. Here is an example

(3)
```
List of attributes for the <moveLink> tag
<!ELEMENT moveLink EMPTY >
<!ATTLIST moveLink id ID prefix="mvl" #REQUIRED >
<!ATTLIST moveLink trigger IDRef #IMPLIED >
<!ATTLIST moveLink source IDRef #IMPLIED >
<!ATTLIST moveLink goal IDRef #IMPLIED >
<!ATTLIST moveLink midPoint IDRefs #IMPLIED >
<!ATTLIST moveLink mover IDRef #IMPLIED >
<!ATTLIST moveLink ground IDRef #IMPLIED >
<!ATTLIST moveLink goalReached ( yes | no | uncertain ) #IMPLIED
>
<!ATTLIST moveLink pathID IDRef #IMPLIED >
<!ATTLIST moveLink motionSignalID IDRef #IMPLIED >
<!ATTLIST moveLink comment CDATA #IMPLIED >
```

For illustration, consider the following partially inline annotated dataset, where each of the markables is tagged with its entity type:[25]

(4) Dataset:
Mia$_{se1}$ *lives*$_{e1}$ *near*$_{ss1}$ *Harvard*$_{pl1}$ *in*$_{ss2}$ *Cambridge*$_{pl2}$*, but works*$_{e2}$ *at*$_{ss3}$ *[Boston College]*$_{pl3}$ *in*$_{ss4}$ *the [Chestnut Hill section]*$_{pl4}$ *of*$_{ss5}$ *Newton*$_{pl5}$ *[east of]*$_{ss6}$ *Boston*$_{pl6}$*. She*$_{se2}$ *crosses*$_{m1}$ $_{pl7}$*[the Charles River]*$_{pa1}$ *and sometimes takes*$_{m2}$ *I-90*$_{pa2}$*, driving*$_{m3}$ *eastward*$_{ss7}$ *[around 8 miles]*$_{mes}$ *to*$_{ms2}$ *[the university]*$_{pl8}$*.*[26]

Unlike TimeML or SpatialML, ISOspace allows a variety of concrete syntaxes based on its abstract syntax that all represent annotations equivalently. Instead of

[25] The noun *Mia* is tagged as se (spatial entity) because it is spatially involved as the figure of the event *lives near Harvard in Cambridge.*

[26] $_{pl7}$ is a non-consuming tag referring to some spot on the Charles River that is crossed.

a commonly adopted XML format, a predicate-logic-like format may be adopted to represent parts of the annotation of Dataset (4) involving ISOspace links, as shown below:

(5) a. *Mia*$_{se1}$ *lives*$_{e1}$ *near*$_{ss1}$ *Harvard*$_{pl1}$ *in*$_{ss2}$ *Cambridge*$_{pl2}$

```
qsLink(qsl1, relType=near, figure=e1, ground=pl1, signal=ss1)
qsLink(qsl2, relType=in, figure=pl1, ground=pl2, signal=ss2)
```

b. *Mia*$_{se1}$ *... works*$_{e2}$ *at*$_{ss3}$ *[Boston College]*$_{pl3}$ *in*$_{ss4}$ *[the Chestnut Hill section]*$_{pl4}$ *of*$_{ss5}$ *Newton*$_{pl5}$ *[east of]*$_{ss6}$ *Boston*$_{pl6}$.

```
oLink(ol1, relType=east, figure=pl5, trigger=ss6,
frameType=absolute, referencePt=east, projective=false/>)
```

c. *She*$_{se2}$ *crosses*$_{m1\ pl1}$ *[the Charles River]*$_{pa1}$ *and sometimes takes*$_{m2}$ *I-90*$_{pa2}$,

```
moveLink(mvl1, trigger=m1, mover=se2, ground=pl1, pathID=pa1)
moveLink(mvl2, trigger=m2, mover=se2, pathID=pa2)
```

d. *driving*$_{m3}$ *eastward*$_{ss1}$ *[around 8 miles]*$_{mes1}$ *to*$_{ms2}$ *[the university]*$_{pl8}$.

```
measure(mes1, value=8, unit=mile, mod=approx)
mLink(ml1, relType=distance, figure=m3, ground=mes1,
val=mes1, ednPoint2=pl8)
moveLink(mvl1, trigger=m3, mover=se2, goal=pl8, motionSignalID=ms2)[27]
```

See Part II, III.g.i for a case study of ISOspace annotation.

4.2.3 SemAF-SR for Semantic Roles

Noticeably since Fillmore's seminal paper on *The Case for Case* [34], semantic roles associated with eventualities have become the core of grammatical inquiries, for they capture the basic semantic relations of participation between an eventuality, expressed mostly by a verb, and its arguments as participants. From these inquires several systematic frameworks on semantic roles have resulted particularly for the purpose of constructing lexical resources in language, such as: FrameNet [33], VerbNet [70], LIRICS [94] and [109], EngVallex [26], and PropBank [90].[28] ISO's SemAF-SR [65] is a result of such efforts with its objectives to provide (1) a data category-based structured way of defining semantic roles with an explicit semantics, (2) a pivot representation based on a framework for defining semantics roles that could facilitate mapping between different formalisms, and (3) a set of guidelines for creating new resources that would be immediately interoperable with preexisting resources[29]

The annotation scheme of SemAF-SR is a tuple $<M,B,R,@>$.[30]. M is a set of markable expressions, extents of a text the types of which are delineated by B, a set of basic entities. B consists of sets of two types, eventuality type (B_e) and participant (individual) entity type (B_x). R is a singleton consisting of a link of various role

[27] A new attribute @dir for the direction of a motion may need to be introduced to annotate a markable such as *eastward*.

[28] The informative annex B in SemAF-SR [65] reviews these existing framewokrs in detail.

[29] See [18], page 41.

[30] The specification of the annotation structure here is much simplified, differing from that presented in [19]

types which relates a basic entity in B_e of an eventuality type to an entity in B_x that participates in the eventuality with a particular semantic role in it. @ is a set of required or optional attribute-value assignments to each element in B and R, such as identifiers, targets for the markables or type specifications.

Here is a simple example showing how semantic roles are annotated, as represented in XML:[31]

(6) a. Text: `The soprano sang an aria very well.`
b. Markables: `The soprano, sang, an aria,`
c. Basic entities, tagged as `<entity>` and `<eventuality>`:

```
<entity xml:id="x1" target="#token1,#token2" entityType="soprano"/>
<entity xml:id="x2" target="#token4,#token5" entityType="aria"/>
<eventuality xml:id="e1" target="#token3" eventFrame="sing.01"
eventualityType="completeiveAccomplishment"/>
```

d. Link, tagged as `<srLink>`:

```
<srLink xml:id=srL1, event="#e1" participant="#x1" semRole="agent"/>
<srLink xml:id=srL2, event="#e1" participant="#x2" semRole="theme"/>
```

The text contains three markable extents, `The soprano`, `sang`, and `an aria`. The first and the third markables are annotated simply as (individual) entities, while the verb is annotated as an eventuality. Then the eventuality is linked with either of the two arguments (Subject and Object) and the type of each link is specified ”agent” and ”theme”, respectively. The first entity (the soprano) is thus interpreted as the agent of the eventuality of singing, and the second entity (an aria) as the theme of the same eventuality.

The informative Annex A of SemAF-SR [65] introduces ISO-semantic roles, mainly based on LIRICS. Table A.2 in the same Annex lists the definitions of the LIRICS semantic roles in the form of ISO data categories. It then relates the semantic roles of LIRICS to those of other frameworks, VerbNet, PropBank, FrameNet, and EngVallex. Likewise, Clause 8 provides guidelines for developing new semantic role frameworks for various languages and domains (Clause 8.1), while showing how to map VerbNet to LIRICS (Clause 8.2).

4.2.4 ISO SemAF-DA: Dialogue Act Annotation

A dialogue act is a unit in the description of communicative behavior. Semantically, these units correspond to changes that the speaker intends to bring about in the information state of an addressee. A dialogue act has two main components: a *communicative function* and a *semantic content*. The communicative function specifies how the semantic content changes the information state of an addressee who understands the speaker’s communicative behavior. In the ISO standard for dialogue act annotation (ISO 24617-2:2012), communicative functions may be qualified in several respects, such as sentiment and certainty; moreover, a dialogue act may

[31] See Annex C.3 Concrete syntax, SemAF-SR [65].

have various kinds of relations to other dialogue acts, which further contribute to its meaning.

Dialogue act annotation is is the marking up of a spoken, written, or multimodal dialogue with information about the dialogue acts that it contains; in the annotation schemes that existed prior to the establishment of ISO 24617-2 and its predecessor DIT^{++} (such as DAMSL, [1]; MRDA [28] ; HCRC Map Task [23]; and COCONUT citeDiEugenioetal98), this annotation was limited to marking up stretches of dialogue with communicative function labels. The ISO annotation scheme, which was developed by an international group of experts inherited the content and structure of the inventory of communicative functions from the DIT^{++} annotation scheme (Bunt, 2009), which provides a solid theoretical and empirical basis. The structure reflects the view that a stretch of communicative behavior may be *multifunctional*, i.e. may correspond to more than one dialogue act. The scheme has therefore been designed to support 'multidimensional' annotation, but as opposed to DAMSL and other annotation schemes, the DIT^{++} and ISO schemes make the notion of multidimensionality precise by providing an explicit definition of 'dimension'.

The ISO 24617-2 annotation scheme has the following notable features:

1. Multidimensional annotation is based on the definition of nine dimensions of interaction, which are distinguished on empirical and theoretical grounds.
2. Communicative functions are either *dimension-specific* and can only be used only in one particular dimension (like Take Turn), or *general-purpose* and can be used in any dimension, like Question, Inform, and Instruct.
3. Dialogue act annotations attach to 'functional segments', defined as minimal stretches of behavior that have one or more communicative functions.
4. 'Multidimensional segmentation' is used: dialogue is segmented in multiple ways, with functional segments for each dimension. A segment carrying a feedback function may for instance overlap with a segment that carries a task-related function.
5. 'Function qualifiers' are defined for expressing that a dialogue act is performed conditionally, with uncertainty, or with a certain sentiment.
6. Functional and feedback dependence relations are defined which relate a dialogue act to units earlier in a dialogue, e.g. for indicating which question is answered by a given answer, or which utterance the speaker is providing feedback about.
7. A markup language is defined, the Dialogue Act Markup Language (DiAML), with a 3-part definition: (1) an abstract syntax, which specifies the possible annotation structures in set-theoretical terms; (2) a semantics which specifies the interpretation of the structures defined by the abstract syntax; (3) a concrete syntax which defines an XML representation of annotation structures.

Dimensions. Utterances in dialogue often have more than one communicative function, as several authors have observed: [2, 12, 13, 96, 117] Dialogue participants share information not only about the task or activity that they pursue, but also about the processing of each other's messages, about the allocation of turns, about contact and attention, and about various other aspects of the interaction. They therefore perform communicative activities such as giving and eliciting feedback, taking turns, stalling for time, establishing contact, and showing attention; moreover, they often perform more than one of these activities at the same time. The term *dimension* refers to these various types of communicative activity or to the types of information that they are concerned with. Supported by an analysis of 18 existing annotation schemes [93] the following nine dimensions are defined:

1. Task: dialogue acts that move the task or activity forward which motivates the dialogue;
2-3. Auto- and Allo-Feedback; dialogue acts providing or eliciting information about the processing of previous utterances by the current speaker or by the current addressee, respectively;
4. Turn Management: activities for obtaining, keeping, releasing, or assigning the right to speak;
5. Time Management: acts for managing the use of time in the interaction;
6. Discourse Structuring: dialogue acts dealing with topic management, opening and closing (sub-)dialogues, or otherwise structuring the dialogue;
7-8. Own- and Partner Communication Management: actions by the speaker to edit his current contribution or to edit (corrupting or completing) a contribution of another current speaker, respectively;
9. Social Obligations Management: dialogue acts for dealing with social conventions such as greeting, introducing oneself, apologizing, and thanking.

Some communicative functions are specific for a particular dimension; for instance *Turn Accept* and *Turn Release* are specific for turn management.

Multidimensional segmentation. Spoken dialogues are traditionally segmented into *turns*, defined as stretches of communicative behavior produced by one speaker, bounded by periods of inactivity of that speaker. However, turns may contain sequences of several dialogue acts. Dialogue act annotation can be done more accurately by using smaller ’functional segments’, defined as the *minimal* stretches of communicative behavior that have a communicative function. Functional segments are mostly shorter than turns but may also stretch over more than one turn, may be discontinuous, may overlap, and may contain parts contributed by different speakers.

Qualifiers. The function qualifiers defined in ISO 24617-2 are applicable to the general-purpose communicative functions (GPFs). Sentiment qualifiers are applicable to any GPF; conditionality qualifiers are applicable to the ‘action-discussion functions’ among the GPFs (such as Promise, Offer, Suggestion, Accept Request, etc.); and certainty qualifiers are applicable to the ‘information-providing’ functions’ GPFs (Inform, Agreement, Disagreement, Correction, Answer, Confirmation, Disconfirmation).

Relations Between Dialogue Acts. In a coherent dialogue the contributions are connected by various relations. *Rhetorical relations*, which have been studied extensively for written texts, also occur in spoken dialogue. Dialogue acts that are responsive in nature, such as Answer, Confirmation, Agreement, Accept Apology, and Decline Offer, have a semantic content that depends crucially on the content of the dialogue act that they respond to (and are often expressed by utterances that by themselves have little or no semantic content, such as *"Yes"* and *"OK"*). *Functional dependence relations* connect occurrences of such dialogue acts to their 'antecedent' and correspond to links in the ISO scheme for marking up a functional segment not only as expressing an answer, for example, but also indicating which question is being answered. Similarly, the semantic content of a feedback act depends on the utterance(s) that the feedback is about. Feedback acts often refer to the immediately preceding utterance, but can also refer farther back and to more than one utterance (Petukhova, Prévot & Bunt, 2011). [95] The ISO 24617-2 annotation scheme includes links to mark up these *'feedback dependence relations'* between feedback acts and the utterances that form their scope of reference.

4.3 ISOcat

See also Chapter III, Section 5.3, in this volume.

ISOcat is not an annotation scheme *per se*, but rather it is a large, web-based reference repository of (mostly) linguistic terminology that provides human-readable descriptions of the meaning of terms used in language resources, such as *grammaticalNumber, gender, case*. ISOcat is often used as a glossary in which users can look up the meaning of a term occurring in a language resource by consulting its ISOcat entry, but for the purposes of linguistic annotation it is a means to achieve *semantic interoperability* [47] among language resources by enabling different annotations to reference the same definition and thus indicate that they have the same meaning. Prior to the development of ISOcat there was no explicit and verifiable means to ensure that the definition of a linguistic category were identical; to address this, ISOcat and repositories like it facilitate semantic interoperability by providing unique URIs for linguistic terms, to which annotations can refer via hyperlinks.

Unlike the ISO standards described in the previous sections, ISOcat was not developed within ISO TC37 SC4, and in fact grew out of ISO work that predated the establishment of ISO TC37 SC 4. ISOcat has its roots in the late 1990s when ISO TC37 (Terminology and Other Language and Content Resources) developed the standard ISO 12620:1999 Data Categories, which provided a paper list of categories originally intended for use by the terminology community [10]. ISO 12620:2009 (Data Category Registry) is a successor of this standard, designed to overcome the limitations of the earlier paper-based standard, in particular regarding extensibility with new data categories. In 2004, a proposal for a registry accommodating the needs of not only the terminology community but also the community of users involved in linguistic annotation of language resources was proposed [53]. The resulting effort was ISOcat, which implemented ISO 12620:2009 as an online repository that was accessible and extensible with new data categories by the community. ISOcat entries comply with the data model defined in the standard, which specifies mandatory information types such as a unique administrative identifier (e.g., partOfSpeech) and a unique and persistent identifier (PID, e.g., http://www.isocat.org/datcat/DC-396) which can be used in automatic processing and annotation, in order to link to ISOcat entries.

As an example, consider the MASC corpus annotated with WordNet [31] senses.[32] By establishing a trivial linking of the WordNet senses to their ISO-LMF compliant lexical entries in a standardized resource such as UBY [42, 29] (based on the WordNet sense keys), the MASC corpus is enriched by further lexical annotations on the sense level, many of which contain terms defined in ISOcat. For instance, verb senses can be enriched by an annotation indicating particular lexical-syntactic properties, such as subject-control[33] or object-control[34]. This is achieved by following the links from WordNet senses to VerbNet [69] senses given in UBY. The reference to the definition of subject-control and object-control in ISOcat makes the meaning of the new annotation transparent and ensures that humans (and also applications built by humans) interpret these annotations in the right way.

4.3.1 Evolution of ISOcat

In the beginning, ISOcat took an open, community driven approach and allowed everybody to sign up, create data categories, and thus extend the repository. Users were allowed to assign their data categories to so-called Thematic Domains, such as Morphosyntax, Syntax, Semantic Content Representation or Lexical Resources, etc. Users were also able to group data categories, including self-created ones, into a Data Category Selection. Data Category Selections can be made publicly available, in order to allow for linking to particular data categories defined within them.

Understandably, from the beginning Data Category Selections tended to be created for specific projects or resources, e.g., large projects like RELISH[35] and CLARIN[36] as well as resources such as the partOfSpeech tagset STTS[37] and the lexical resource UBY[38], which tended to limit generality and lead to the creation of multiple entries for the same concept. More generally, the openness of the repository turned out to be a major reason for the proliferation of data categories, which degraded the usability of ISOcat as a source of a common and unique terminology.

ISOcat's usability issues became especially problematic in the large EU CLARIN project, as summarized in [10]. First, as mentioned above, there are no mechanisms that prevent the creation of new data categories that are almost equivalent to existing ones [119]. Such near-equivalents were introduced by bulk imports of whole sets of data categories, such as the STTS tagsets. This made the selection of ap-

[32] See Part II, III.a

[33] http://www.isocat.org/datcat/DC-4187

[34] http://www.isocat.org/datcat/DC-4189

[35] http://tla.mpi.nl/relish/

[36] http://www.clarin.eu

[37] http://www.isocat.org/rest/dcs/376

[38] http://www.isocat.org/rest/dcs/484

propriate data categories among existing ones for a particular language resource a tedious task. Second, no standardized data categories were available, and at the same time, the procedures around standardizing data categories turned out to be impractical; alternative approaches such as mechanisms for community control and approval of data categories were not in place. Standardized data categories are important for resources that comply with other ISO standards, such as the Lexical Markup Framework (ISO 24613, 2008) and the Linguistic Annotation Framework (ISO 24612, 2012), both of which require or recommend reference to ISOcat data categories. Also, standardized data categories can be considered as stable and therefore contribute to the sustainability of a language resource. Non-standardized data categories, on the other hand, could in principle be changed at any time by their owners which might also involve changes in their meaning.

FInally, the data model defined by the Data Category Registry standard was perceived as too complex by many users from the language data research community, especially those who are not technically sophisticated. The data model distinguishes three different data category types [120]:

- *Complex Data Categories* have a conceptual value domain. According to the size of the value domain, Complex Data Categories are classified further into *Open Data Categories* (they can take an arbitrary number of values), *Closed Data Categories* (their values can be enumerated) and *Constrained Data Categories* (the number of values is too big in order to be enumerated, but yet constrained).
- *Simple Data Categories* describe values of a Closed Data Category.
- *Container Data Categories* are used to group other data categories together

Moreover, the data model itself was another source of the proliferation of data categories, since the data type of a data category is determined by its use in a particular language resource [119].

In order to address these issues, the ISO TC37 and CLARIN communities recently met and discussed the future of ISOcat as a Data Category Registry. Among the fundamental issues noted were a conflict of interest between the ISO terminology community and the language resource community: the terminology community requires broad definitions that are applicable to as many languages as possible, for terminological consistency. The language resource community, on the other hand, requires definitions that describe term usage in specific contexts, as well as a more rapid and efficient system for achieving agreement with possibly limited scope and feedback, coupled with considerable human coordination. It was also noted that the creation of nearly equivalent data categories is unavoidable, given the differences in theoretical perspective as well as the broad range of applications for which the categories are used.

As a result of this meeting, the two communities ultimately split the ISOcat work into two efforts. It was agreed that CLARIN would simplify the data model by focusing on the semantics of a data category, which is described by a single data type, the concept. The data categories relevant to CLARIN have now been transformed to concepts and migrated to the new *CLARIN Data Concept Registry*.[39]

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

This new registry is a closed repository where only the national CLARIN Concept Registry coordinators are able to input and edit concepts; they are also the only persons eligible for marking concepts as "accepted" as an official CLARIN standard or recommendation. ISO TC37, on the other hand, continues to be responsible for http://www.isocat.org/ and plans to launch a new implementation of the Data Category Registry in collaboration with a commercial provider of terminology management software. How exactly the new version of the Data Category Registry will address the issues described above is currently being discussed as part of the transition process.

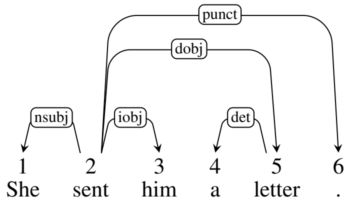
4.4 Concluding Remarks

Within a decade or so after it was established, ISO SC 4 published a dozen ISO standards for language resources, and it continues to produce standards for new phenomena. However, more work is needed to achieve better integration and interoperability among the published and developing standards. [78] make several recommendations based on LAF (2012) [64] and TEI Guidelines P5 [113] that if adopted, will make a significant move toward development of a single, unified format for language resource annotation and representation that will serve sustainable use and applications.

Whatever its future, the creation of ISOcat has served as a landmark exercise in the effort to achieve semantic interoperability among linguistically-annotated resources, from which much has been learned to guide future development. The underlying notion of linkage via web-based reference to achieve semantic consistency is fundamental to the Linked Data (Semantic Web) model, which is taking on increasing centrality in the language resources domain. However, despite the promise of the Linked Data model, the experience of ISOcat dramatically underscores the challenges of defining and modeling linguistic concepts that remain.

5 *De Facto* Standards: CoNLL and Dependency Annotations

After being a rather marginal phenomenon in natural language processing only one decade ago, dependency parsing has evolved into one of the mainstream approaches to syntactic parsing [73]. The dramatic increase in popularity and usage has naturally led to a need for standardization with respect to input and output formats for parsing, as well as for interchange of treebank data with dependency annotation. Given that a dependency tree can be specified simply by assigning to each word a syntactic head (another word in the sentence) and a dependency relation, most parsers and treebanks use a representation where each word in a sentence has two essential attributes: a head index and a dependency label. This is illustrated in Figure 5, which shows a dependency tree (left) and its representation in the Malt-TAB



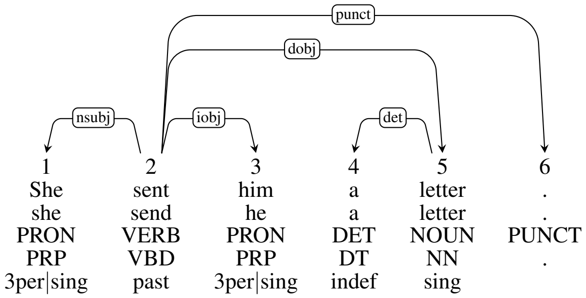
```
1 She    2 nsubj
2 sent   0 root
3 him    2 iobj
4 a      5 det
5 letter 2 dobj
6 .      2 punct
```

Fig. 5 The Malt-TAB format.

format used in the first release of MaltParser [89], where each line represents a word token with four attributes: index, word form, head index, dependency label.

An important milestone in the development of dependency parsing was the CoNLL-X shared task on dependency parsing [11], which involved data sets for 13 different languages. To make it possible to train and evaluate a single parser on all languages, the shared task organizers had to devise a new format that was expressive enough to capture the annotation in the 13 native annotation formats. For this purpose, they generalized the simple Malt-TAB format by adding more attributes to each word token and created the CoNLL-X format, defined as follows[11]:

> All the sentences are in one text file and they are separated by a blank line after each sentence. A sentence consists of one or more tokens. Each token is represented on one line, consisting of 10 fields. Fields are separated from each other by a TAB. The 10 fields are:
>
> 1) ID: Token counter, starting at 1 for each new sentence.
>
> 2) FORM: Word form or punctuation symbol. [. . .]
>
> 3) LEMMA: Lemma or stem (depending on the particular treebank) of word form, or an underscore if not available. [. . .]
>
> 4) CPOSTAG: Coarse-grained part-of-speech tag, where the tagset depends on the treebank.
>
> 5) POSTAG: Fine-grained part-of-speech tag, where the tagset depends on the treebank. It is identical to the CPOSTAG value if no POSTAG is available from the original treebank.
>
> 6) FEATS: Unordered set of syntactic and/or morphological features (depending on the particular treebank), or an underscore if not available. Set members are separated by a vertical bar (|).
>
> 7) HEAD: Head of the current token, which is either a value of ID, or zero (0) if the token links to the virtual root node of the sentence. Note that depending on the original treebank annotation, there may be multiple tokens with a HEAD value of zero.
>
> 8) DEPREL: Dependency relation to the HEAD. The set of dependency relations depends on the particular treebank. The dependency relation of a token with HEAD=0 may be meaningful or simply ROOT (also depending on the treebank).


punct
dobj
nsubj
iobj
det
1 She she PRON PRP 3per|sing
2 sent send VERB VBD past
3 him he PRON PRP 3per|sing
4 a a DET DT indef
5 letter letter NOUN NN sing
6 . . PUNCT .

1	She	she	PRON	PRP	3per|sing	2	nsubj	_	_
2	sent	send	VERB	VBD	past	0	root	_	_
3	him	he	PRON	PRP	3per|sing	2	iobj	_	_
4	a	a	DET	DT	indef	5	det	_	_
5	letter	letter	NOUN	NN	sing	2	dobj	_	_
6	.	.	PUNCT	.	_	2	punct	_	_

Fig. 6 The CoNLL-X format.

9) PHEAD: Projective head of current token, which is either a value of ID or zero (0), or an underscore if not available. The dependency structure resulting from the PHEAD column is guaranteed to be projective (but is not available for all data sets), whereas the structure resulting from the HEAD column will be non-projective for some sentences of some languages (but is always available).

10) PDEPREL: Dependency relation to the PHEAD, or an underscore if not available.

The CoNLL-X format, which is illustrated in Figure 6 for the same sentence as in Figure 5, quickly became the de facto standard for dependency parsing, and all available dependency parsers today accept input and output in this format. The format is also widely used for dependency treebanks, in particular as an interchange format, but there are also many treebanks that use their own format to overcome the limitation that the CoNLL-X format is restricted to dependency trees and does not support multiheaded structures.

It is important to note that the CoNLL-X format only standardizes the encoding of dependency structures and does not have anything to say about which labels to use or about the criteria for determinining syntactic heads in different languages. Until recently, it has therefore been the case that virtually every dependency treebank has its own unique annotation scheme, which is then inherited by statistical parsers trained on a given treebank. As a consequence, it has been very difficult to compare parsing results across languages and to properly evaluate systems for cross-lingual learning in the domain of dependency parsing [88, 84]. To overcome these difficulties, there have been a number of recent initiatives to create a standard for cross-linguistically consistent dependency annotation [122, 83, 118, 80]. Several of these initiatives have now been merged into the Universal Dependencies (UD) project, which released the first guidelines for cross-linguistically consistent annotation in October 2014 and the first set of ten treebanks in January 2015.[40] The UD consortium has also proposed a revised version of the CoNLL-X format called CoNLL-U. The main difference between CoNLL-U and CoNLL-X, except for the use of universal part-of-speech tags, morphosyntactic features and dependency labels, is that CoNLL-U supports the representation of multi-headed structures as well as two levels of word segmentation.

In addition to standardizing the format, the UD guidelines also standardize the linguistic content of the annotation, by providing three sets of linguistic categories:

1. Part-of-speech tags: This is a revised and extended version of the Google Universal Part-of-Speech Tagset [92] containing 17 tags. These tags must be used without exception (although it is conceivable that some languages do not use all of them).
2. Morphological features: This is an inventory of features based on Interset [121], which is a consensus standard based on a large number of existing tagsets, previously used as an interlingua for tagset conversion. Each language uses a selection of these features, but it is also possible to define language-specific features if needed.
3. Dependency relations: This is a set of 42 basic grammatical functions based on the Universal Stanford Dependencies [80], which in turn is an adaptation of the original Stanford Dependencies for English [81, 82]. These labels must be used without exception, but it is possible to define language-specific subtypes of the universal relations.

The experience from the first UD release shows that, even if the categories proposed are adequate and sufficient (given the possibility of adding language-specific features and dependency subtypes), more detailed guidelines for the *use* of different categories are needed. For instance, the distinction between determiners and pronouns is drawn differently in different traditions, so just providing two universal part-of-speech tags DET and PRON does not necessarily lead to a cross-linguistically consistent annotation. Similarly, at the syntactic level, there is a need for more detailed guidelines at the level of grammatical constructions, as opposed to individual grammatical relations. Adding these guidelines to ensure consistent application of different categories is probably a necessary step in order for the standard to gain wide acceptance.

6 Standards for Spoken Language Data

The terms “spoken language” and “speech” characterize domains of research and application in several disciplines, from phonetics, language teaching and documentary field linguistics through sociology, psychology and speech pathology, some of which have become associated with the meta-discipline of digital humanities, to

[40] See http://universaldependencies.github.io/docs/

computational models of components of spoken language and speech technology, each with their sub-disciplines, and each with their theories, models, terminologies and de facto or institutional standards for best practices. The problems which arise from this multidisciplinary diversity are considerable: the institutional standard ISO 639-3 codes for the identification of languages are a starting point for shared information, but are still rarely applied, even publications in linguistics, phonetics and the speech technologies. There are few institutional standards, when the spoken language domain is seen as a whole, and the field is largely in flux, but there are de facto standards and trends.

The present outline of standards for spoken language will first characterize the domains and properties of spoken language as a basis for further discussion, then outline standards developments for basic resources shared by a number of disciplines, such as transcription and speech signal annotation, and for the development and quality control of spoken language resources. The speech technologies of automatic speech recognition, text-to-speech synthesis, and language and speaker identification are not treated in detail here; rather, the focus is more on linguistic and phonetic requirements and standards which are relevant to computational scenarios.

6.1 Domains of spoken language: standards versus diversity

Spoken language domains cover a wide range of communication styles, genres and scenarios: communication styles (from intimate through informal to formal), genres (e.g. interview, joke, narrative, public speech, sermon) and scenarios (monologue, face-to-face, audio and video phone, one-way mass media). Historically and in child language development, speech precedes written language, and may itself be predated by gestural communication [85, 37, 106]. Indeed, speech is a form of gestural communication transduced into the acoustic medium, just as writing, at the physical level of manuscript production, is a transduction of gesture into visible inscriptions. Each modality has different consequences for communication speed, support by memory and cognitive processes, distance coverage in space and durability in time.

The speech-text modality differences also have practical, scientific, ethical and forensic consequences [39, 38, 4]. Speakers, unlike writers, are often instantly recognisable within fractions of a second, yet their speech is not durable unless recorded on a technical medium. In many scenarios speech is temporally and locally coextensive with gestural and tactile communication modalities; in other scenarios the modalities are separated (e.g. in visually or acoustically challenging situations), or the speech setting is subject to dislocation in speech at a distance (teleglossia, e.g. in telephony and visual conferencing) and distemporality (e.g. in writing). Speech is increasingly seen as multimodal, together with gestural and tactile interaction, and multimodal speech in technical communication has become a major subdomain [86].

The speech-only communication domain is typically found in the oral societies which remain in some parts of Africa, South America and South East Asia, studied by field linguists, ethnologists and anthropologists, often in cooperation with other disciplines such as musicology [4]. A large part of daily communication in industrially and economically developed societies is substantially similar, though complemented by complex varieties of communication in technically transmitted media, from writing, whether with pencil, stylus, phone or PC, or multimodal internet telephony. Influential scientific conference series such as *Interspeech* (mainly speech engineering), the *International Congress of Phonetic Sciences* (mainly the physical modality aspects of spoken language) and the *Language Resources and Evaluation Conference* (LREC) bear witness to the diversity not only of the domain but of methodologies, and many conferences and journals in other disciplines give implicit or explicit coverage to spoken language.

There is thus no single spoken language research, development and application community, as the present discussion shows, and consequently *de facto* standards for data, tools and information interchange have developed differently in the different communities, and sometimes even basics like phonetic transcription are not uniformly practiced. Another factor which militates against the development of comprehensive sets of standards is the complexity of the field and the disparity of topics and R&D interests:

1. Spoken and written language differ not only in the phonetic and prosodic modalities and their levels of representation, but also in the lexicon (e.g. levels of style; hesitation phenomena and other discourse particles), the grammar (e.g. levels of style, rarity of centre-embedding except in formal styles, disfluency handling strategies), and at discourse levels (e.g. turn-taking, turn overlap).
2. In crucial respects the semantics and pragmatics of spoken and written language differ (e.g. in deictic and utterance act properties).
3. Spoken language occurs concurrently and coordinated with visible gestural and postural communication (for a recent account, cf. [106]) and is itself gesture.
4. Quality criteria, size, accessibility, ethical and legal status of spoken and written data differ.
5. The tools for processing spoken language at the phonetic levels (production, transmission, reception) are specialized and only comparable with the tools for studying written language in terms of manual gesture in handwriting production, typing and touchscreen input, and with the optical character and layout recognition of handwritten and electronically formatted manuscripts and touchscreen gesture signals.

In spite of the speech-text differences, lexical properties of spoken language can in general be catered for by existing lexicographic conventions, and grammatical properties by existing tagset and Treebank conventions, except for the lattices used to represent word hypotheses in speech recognition or turn overlap in discourse analysis. For an ISO standard for dialogue act categories cf. [17].

Spoken language has specific characteristics at all ranks of linguistic description from speech sounds through phonemes, morphemes, words, phrases and sentences, to utterances and discourse. Compared with constituents of text, units at each of these ranks have their own properties of interpretation, both semantic and phonetic.

Semantic interpretation ranges from bare contrastivity of phonemes, through morphemes and words as predicates and operators, to sentences as propositions, texts as argumentation and discourse as negotiation. Phonetic interpretation ranges from sequential segmental consonantal and vocalic patterns and their hierarchical organization in syllables and larger groups to concurrent prosodic (suprasegmental) rhythmic and melodic features such as phonemic tone, morphemic tone, accentuation, and higher ranking intonational and rhythmic patterning at sentence and discourse levels.

While there are institutional standards for transliteration (i.e. the conversion of one system of writing into another, e.g. ISO 9 for Cyrillic or ISO 15919 for Indic scripts), there is currently no ISO standard for phonetic and phonemic transcription. However, professional curating of standardization in the phonetic and phonemic representation of language is administered by the International Phonetic Association, and the alphabet, including diacritics, has a complete Unicode encoding.

There is one outstanding set of professional *de facto* standards which is used in all of the spoken language communities, from linguistic theory and fieldwork research to applications in language teaching and speech pathology to the spoken language technologies: the IPA[41], the IPA character coding according to the Unicode standard, and the formulation of descriptive rules for phonetic processes, such as assimilation, based on the IPA. The IPA is an empirical standard, and has evolved as empirical knowledge has developed, with extensions for specialized purposes such as speech pathology. The IPA was originally conceived as an alphabet which can represent all speech sounds which are contrastively phonemic in all languages of the world. The current understanding of the IPA is more phonetic, and the alphabet is intended to represent all identifiable speech sounds, whether contrastive or not. For the representation of phonemes in languages with less common IPA characters, very often these are substituted with no loss of information (if properly defined) by more common characters which are easier to type.

The *International Phonetic Alphabet* (IPA) has been curated since 1886 by the main professional body in phonetics, the International Phonetic Association[42] (also IPA). The segmental categories, characters and glyph sets of the IPA are widely accepted as a standard point of reference, but there are many specific application-oriented variant alphabets. Divergent segmental transcription conventions are used in the historical philologies and in anthropological language studies. Extensions of the IPA have been proposed for specialized use cases, for example in speech pathology [114]

Although the IPA is fully specified in Unicode, IPA codes are scattered over a number of code blocks, presumably for the sake of space economy, where particular symbols are used in the official orthographies of various languages (e.g. "θ" in the Greek block, or "ð" in the Latin-1 blocks). This dispersion of characters frequently leads to uncertainty and inconsistency in use by picking similar but differently coded characters. The lack of a coherent use case semantics for code block allocations in

[41] https://www.internationalphoneticassociation.org/content/ipa-chart

[42] https://www.internationalphoneticassociation.org/

Unicode in order to overcome this dispersion property has received some criticism (Hughes et al. 2006). Although many fonts now implement the IPA Unicode characters, many still do not or are proprietary. For this reason, in linguistics the Gentium[43] font of the SIL is frequently used and often recommended for publications.

In the speech technologies a number of keyboard friendly encodings of the IPA have been developed, the most widely used being the *SAMPA/X-SAMPA (Speech Assessment Methodologies Phonetic Alphabet*, the “X” stands for “eXtended”; cf. Gibbon et al. 2000). The SAMPA/X-SAMPA coding was originally developed in a EU project as an international consensus of speech engineers and phoneticians for easy information interchange. The SAMPA/X-SAMPA alphabet, being a one-to-one encoding of the IPA, is widely (though not exclusively) used internally in system development in preference to Unicode (ISO 10646) for practical reasons, mainly for being human readable and keyboard friendly and not requiring UTF-*n* codecs. Another reason is that Unicode development focuses on rendering on print output devices rather than on efficiency in character input, and print is not always relevant in spoken language computing contexts.

Symbol sets for prosodic transcription are characterized by much greater diversity, which starts at the level of phonemic tone, with numbers 1 to 5 for Mandarin tones, through the accent diacritics , ‘, în Africanist linguistic usage for high, low, high-low etc. tones (and the same diacritics for rising, falling, rising-falling, etc. in intonational pitch contours), to the IPA symbols for tones. These prosodic transcription notations represent categories. In experimental phonetics and speech technology, a categorial system, *ToBI (Tones and Break Indices)* has become widely used, though it has limitations for tone languages on the one hand and discourse intonation contours on the other. A relational transcription, e.g. IntSint (mainly applied to speech synthesis; [44]), which represents pitch ranges and pitch changes within a coherent acoustic model, has a different semantics in the phonetic domain from the categorial systems. There are many other systems of prosody transcription besides these, some of which are based on explicit models of speech production or perception, which will chiefly interest specialists in phonetics, psychoacoustics and speech technology.

As with many standards, there are limitations on practical use cases for the IPA. The IPA standard is particularly relevant for the display of IPA characters on screen or printed page. Although IPA is easy to write by hand, there is currently no accepted standard for keyboard input. The main methods are:

1. ad hoc keyboard short-cut tools for IPA subsets,
2. internet character selection tables, conversion tools and online keyboards,
3. menu based character tables in word processors.

Perhaps the most ergonomic method for manual input to use the SAMPA keyboard-friendly encoding, and to copy and paste using a converter from SAMPA/X-SAMPA (ASCII) to IPA (Unicode). Tools for all of these methods are easily found on the internet; no specific addresses are given since fluctuation is high. An optimal solution

[43] http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=gentium

would be a touchscreen display based on the standard IPA chart, either on-screen or as an an "IPA mouse". Currently there is much discussion on these unresolved issues and the challenge remains open.

In computational linguistic and software development environments, the internal representation of IPA characters as Unicode or SAMPA or in other internal codings is not an issue as any of these can be easily handled with a conversion table, as the encodings are biunique; the issues are concerned with user interfaces. Very common in a number of technological contexts are also lexicon and rule-based grapheme-phoneme converters for specific languages. The current standard format for text data storage, including IPA, is to use XML with Unicode entities, as in other domains, and the integration of spoken language information into XML formats on this basis is unproblematic.

6.2 Spoken language resources: annotation standards

The structural and functional markup notations of Natural Language Processing, such as part of speech or dialogue act tagging [17] are frequently referred to as "annotation. The term "annotation" has a somewhat different meaning in the spoken language technologies and in empirical studies of spoken discourse, where it refers to the assignment of time-stamps aligned with the speech signal to transcription symbols or to structural and functional markup.

Before annotation types which also apply to writing (part of speech tagging, tree structure annotation, etc.) are applied in the spoken language domain, modality specific annotation is required. The speech signal is recorded digitally and annotated manually, semi-automatically or automatically using appropriate tools, by assigning transcription labels to time-stamps aligned with the signal. A distinction is commonly made between segmentation, i.e. the assignment of boundary time-stamps to speech signals, and labelling or annotation, i.e. the assignment of a transcription symbol to interval or point time-stamps. The distinction is parallel to the traditional "segmentation and classification" procedures in linguistic data treatment. There are currently no general institutional standards for speech signal annotation, but a number of widely used de facto standards for specific purposes have emerged.

Formal definitions for annotation systems were given by [6] and applied to annotation by [36]. More general *annotation graphs*, applicable to both text and speech markup, were defined formally by [7]. Summarizing: A spoken language annotation A has two hierarchical levels:

1. A set of information tiers (vectors, streams) T of labels $L_1, , L_n$, each T representing different information about the speech signal (e.g. phonetic information such as speech sounds, tones, intonation, syllables, words, structural information such as parts of speech or functional information such as discourse functions).
2. Each label L is a pair $< E, S >$ of an event representation E, i.e. a transcription symbol, and a time-stamp S, which is a representation of either an interval I or a

point P. The interval Imay be understood either as a pair of start and end points P_s and P_e, or by a point P_s and a duration D, or a duration D and a point P_e. The point representations are timestamps.

The implementation of annotation data types varies considerably. An early data type was dyadic, a pair of a transcription symbol for an event, paired with a single time-stamp for the interval start (and often system-specific codes, e.g. for color representation in screen visualizations). A constraint on this pair annotation data type is that the speech recording must, in principle, be exhaustively annotated, otherwise interval ends are unspecified. A different dyadic data type is point event and time-stamp, which has a different temporal semantics from the symbol plus interval start time-stamp.

The most common speech annotation implementation is a *triple* consisting of a transcription symbol and two time-stamps, for the start and end of an interval). The triple annotation type permits partial annotation of a speech signal, since each annotation interval is fully specified. A specialized type of triple system is used for diphone-based speech synthesis, where the semantics of the event is different from other systems: the “event” is defined as extending from the temporal centre or acoustically salient peak of one speech sound to the centre or peak of the next. A variant which has been used in speech synthesis has a quadruple format: the label, and three time-stamps for start, centre or peak, and end of the interval.

There are two main use cases for spoken language annotation: first, in speech technology, where annotation is primarily fully automatized and based on machine-learning principles; second, in linguistic phonetics and linguistics from phonology to discourse analysis, where annotation is typically manual, using annotation visualization tools, and annotation mining for descriptive purposes is semi-manual and often spreadsheet based. The following discussion will concentrate on the linguistic use case. There are several high quality and widely used tools available for phonetic annotation, some for transcription alone (e.g. Transcriber), some in a phonetic workbench (e.g. Praat, Wavesurfer, Annotation Pro), and others in a multimodal annotation environment (e.g. Elan, Anvil).

The *de facto* standard annotation tool for linguists and linguistically oriented phoneticians is the Praat phonetic workbench [8], though new annotation tools with enhanced analysis facilities are continually appearing. New developments in providing automatic annotation for linguistic purposes are also appearing, and will lead to the development of new and more efficient workflow practices in this area (e.g. SPPAS [5]).

Non-computationally interested users are usually interested in the visualizations provided by the tools, not the internal and interchange formats used by these tools, and in the manual or automatic methods for deriving linguistic and phonetic descriptions from the annotations.

Currently the most common formats for information interchange of manual annotations in computational contexts are textual, with either character separated value (CSV) format of an annotation triple ¡label, timestamp, timestamp¿, or the “TextGrid” format developed for the Praat phonetic workbench [8], both dating from pre-XML days. For timestamps, the Praat format uses seconds in a decimal format, while some other formats use milliseconds. The CSV formats can be enhanced *ad hoc* by a metadata header using comment lines. The Praat format has been criticized for not including provision for extensive metadata. The Praat format has each information item on a separate line, and may be represented in a generalized form by the following expression (without regard for line formatting):

$$\texttt{metadata tiercount}_n\ \texttt{(tiername intervalcount}_m\ \texttt{(timestamp}_i\ \texttt{timestamp}_i+1\ \texttt{label)}^m\texttt{)}^n$$

The expression is not strictly a regular expression because of the dependency between the subscript and superscript n and the subscript and superscript m, and the temporal immediate precedence constraint between the subscripts i and i+1. The definition also applies, at this level of generality, to the main features of CSV formats.

So far there is no agreed XML standard for speech annotation, though several tools provide export into XML formats. For general computing and archiving purposes, standard CSV formats with metadata comments, and column and row headers are at least as perspicuous as the more verbose formats.

For conversion between formats and for speech annotation mining and manipulation many tools are available (e.g. the online Time Group Analyzer[44] (TGA) [71], Python modules (e.g. TextGrid Tools[45] [21], and many Praat scripting applications[46].

6.3 *Outlook: technology, quality assessment and standards convergence*

The major venues for the dissemination of results in standards development for spoken language systems are the series *Interspeech* and *LREC*, while the *COCOSDA (International Coordinating Committee for Speech Databases and Assessment)* initiative, particularly the annual conferences of its East Asian Branch, *Oriental COCOSDA*, plays a role in focussing attention on standards for resources and system development in the speech technologies.

For practical purposes, different speech technologies may be distinguished, for which different standardization requirements are needed, the main technologies being automatic speech recognition (ASR), text-to-speech synthesis (TTS), language identification and speaker identification. There are several ISO and national standards which refer to quality control aspects of these systems, particularly in safety relevant environments, such as the audibility of announcements in acoustically challenging scenarios such as underground train stations and on speech in telecommunications transmission systems, such as GSM encoding, and other acoustic encodings

[44] http://wwwhomes.uni-bielefeld.de/gibbon/TGA/

[45] https://github.com/hbuschme/TextGridTools/

[46] http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html

such as WAV, WMA and MP3. Reference may be made to the standard handbooks for information on relevant standards for technical communication (e.g. [39, 38, 86].

Although the current situation in the field of spoken language resources, in particular databases and tools, is very heterogeneous, there are nevertheless factors which are gradually leading to convergence in the interests of resource quality and information interchange, the main pressures predictably being the need for reusability of data and the interoperability of tools.

There several national and international centers concerned with the assessment of the quality of speech databases, mainly in the context of data exchange for speech technology research and development (e.g. ELRA/ELDA, Paris), and there is a great deal of ongoing work on inter-transcriber and inter-annotator reliability and consistency. The work on consistency parallels, to a large extent, work on text markup reliability and consistency assessment, except that annotation also has the property of being time-aligned, so that variations in the centi-second region need to be assessed as similar or dissimilar. The studies by [9] and [112] of inter-annotator agreement for two prosodic annotation systems demonstrate current evaluation methods.

The second major influence on convergence towards shared standards is the use of de facto standard interoperable software tools whose formats and visualization provide benchmarks for the development of future resources.

There are signs in current internet discussion, conference contributions and institutional standardization initiatives that collaboratively motivated standards for spoken language are emerging in the following areas:

1. Transcription: IPA, in spite of small divergence for specific application areas, as a durable transcription standard.
2. *De facto* “favorite” standards for annotation tools and formats, e.g. Praat, though new tools for other use cases and with more facilities are continually emerging.
3. Standards for spoken language database quality assessment in terms of comparison algorithms for different domains.

7 Toward Linked Data: NLP Interchange Format (NIF)

An important prospect for improving the quality of linguistic annotations is the availability of large quantities of qualitative background knowledge on the currently emerging Web of Linked Data [3]. Many annotation tasks can greatly benefit from making use of this wealth of knowledge being available on the Web in a structured form as *Linked Open Data* (LOD). The precision and recall of Named Entity Recognition, for example, can be boosted when using background knowledge from DBpedia, Geonames or other LOD sources such as crowdsourced, community-reviewed and timely-updated gazetteers. Of course, the use of gazetteers is a common practice in NLP. However, before the arrival of large amounts of Linked Open Data their creation and maintenance in particular for multi-domain NLP applications was often impractical.

The use of LOD background knowledge in NLP applications poses some particular challenges. These include: *identification* – uniquely identifying and reusing identifiers for (parts of) text, entities, relationships, NLP concepts and annotations etc.; *provenance* – tracking the lineage of text and annotations across tools, domains and applications; *semantic alignment* – tackling the semantic heterogeneity of background knowledge as well as concepts used by different NLP tools and tasks.

In order to simplify the combination of tools, improve their interoperability and facilitate the use of Linked Data we developed the *NLP Interchange Format* (NIF), an RDF/OWL-based format that aims to achieve interoperability between *Natural Language Processing* (NLP) tools, language resources and annotations. The NIF specification was released in an initial version 1.0 in November 2011[47] and known implementations for 30 different NLP tools and use cases (e.g. *UIMA*, *Gate's ANNIE* and *DBpedia Spotlight*) exist and a public web demo[48] is available. NIF addresses the annotation interoperability problem on three layers: the *structural*, *conceptual* and *access* layer. NIF uses a Linked Data enabled URI scheme for identifying elements in (hyper-)texts that are described by the *NIF Core Ontology* (structural layer) and a selection of ontologies for describing common NLP terms and concepts (conceptual layer). NIF-aware applications produce output adhering to the NIF Core Ontology as REST services (access layer).

7.1 URI Schemes

The idea behind NIF is to allow NLP tools to exchange annotations about text in RDF. Hence, the main prerequisite is that text becomes referenceable by URIs, so that they can be used as resources in RDF statements. In NIF, we distinguish between the *document* d, the *text* t contained in the document and possible *substrings* s_t of this text. Such a substring s_t can also consist of several non-adjacent characters within t, but for the sake of simplicity, we will assume that they are adjacent for this introduction. We call an algorithm to systematically create identifiers for t and s_t a *URI Scheme*. To create URIs, the URI scheme requires a document URI du, a separator sep and the character indices (begin and end index) of s_t in t to uniquely identify the position of the substring. The canonical URI scheme of NIF is based on RFC 5147 [49], which standardizes fragment ids for the text/plain media type. According to RFC 5147, the following URI can address the first occurrence of the substring "Semantic Web" in the text (26610 characters) of the document http://www.w3.org/DesignIssues/LinkedData.html with the separator #: http://www.w3.org/DesignIssues/LinkedData.html#char=717,729 The whole text contained in the document is addressed by "`#char=0,26610`" or just "`#char=0,`". NIF offers several such URI schemes which can be selected accord-

[47] http://nlp2rdf.org/nif-1-0/

[48] http://nlp2rdf.lod2.eu/demo.php

[49] http://tools.ietf.org/html/rfc5147

ing to the requirements of the use case. Their advantages and disadvantages have been investigated in [43] and we will limit ourselves to RFC 5147 in this paper. For practical reasons, the document URI and the separator are henceforth called the `prefix` part of the URI scheme and the remainder (i.e. "`char=717,729`") will be called the `identifier` part. NIF recommends the prefix to end on slash (`/`), hash ("#") or on a query component (e.g. `?nif-id=`). Depending on the scenario, we can choose the prefix in the following manner:

1. WEB ANNOTATION. If we want to annotate a (web) resource, it is straightforward to use the existing document URL as the basis for the prefix and add a hash ("#"). The recommended prefix for the 26610 characters of http://www.w3.org/DesignIssues/LinkedData.html is: http://www.w3.org/DesignIssues/LinkedData.html#
   This works best for plain text files either on the web or on the local file system (`file://`). For demonstration purposes, we minted a URI that contains a plain text extraction (19764 characters) created with 'lynx –dump', which we will use as the prefix for most of our examples: http://persistence.uni-leipzig.org/nlp2rdf/examples/doc/LinkedData.txt# and http://persistence.uni-leipzig.org/nlp2rdf/examples/doc/LinkedData.txt#char=33
   NIF can be used as a true stand-off format linking to external text.
2. WEB SERVICE. If the text is, however, sent around between web services or stored in a triple store, the prefix can be an arbitrarily generated URN[50]. Communication between the NLP tools in NIF is done via RDF and therefore mandates the inclusion of the text in the RDF during the POST or GET request. The main purpose here is to exchange annotations between client and server and the used URIs do not require to resolve to an information resource. NIF requires each web service to have a parameter "prefix" that empowers any client to modify the prefix of the created NIF output. The prefix parameter can be tested at http://demo.nlp2rdf.org/.
3. ANNOTATIONS AS LINKED DATA. For static hosting of annotations as linked data (e.g. for a corpus), the `/` and query component separator is advantageous. Often the basic unit of a corpus are the individual sentences and it makes sense to create individual prefixes on a per sentence basis.

In the following, we show how the relation of document, text and substring can be formalized in RDF and OWL.

7.2 NIF Core Ontology

The NIF Core Ontology[51] provides classes and properties to describe the relations between substrings, text, documents and their URI schemes. The main class in the ontology is `nif:String`, which is the class of all **words over the alphabet of Unicode characters** (sometimes called Σ^*). We built NIF upon the Unicode Nor-

[50] http://tools.ietf.org/html/rfc1737

[51] http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#

malization Form C, as this follows the recommendation of the RDF standard[52] for `rdf:Literal`. Indices are to be counted in code units. Each URI scheme is a subclass of `nif:String` and puts further restrictions over the syntax of the URIs. For example, instances of type `nif:RFC5147String` have to adhere to the NIF URI scheme based on RFC 5147. Users of NIF can create their own URI schemes by subclassing `nif:String` and providing documentation on the Web in the `rdfs:comment` field.

Another important subclass of `nif:String` is the `nif:Context` OWL class. This class is assigned to the whole string of the text (i.e. all characters). The purpose of an individual of this class is special, because the string of this individual is used to calculate the indices for all substrings. Therefore, all substrings have to have a relation `nif:referenceContext` pointing to an instance of `nif:Context`. Furthermore, the datatype property `nif:isString` can be used to include the reference text as a literal within the RDF as is required for the web service scenario. An example of NIF Core can be seen on the top left of **??**.

The NIF ontology[53] is split into three parts: The *terminological model* is lightweight in terms of expressivity and contains the core classes and properties. Overall, it has 125 axioms, 28 classes, 16 data properties and 28 object properties. The *inference model* contains further axioms, which are typically used to infer additional knowledge, such as transitive property axioms. The *validation model* contains axioms, which are usually relevant for consistency checking or constraint validation[54], for instance class disjointness and functional properties. Depending on the use case, the inference and validation model can optionally be loaded. Overall, all three NIF models consist of 177 axioms and can be expressed in the description logic $\mathcal{SHIF}(\mathcal{D})$ with exponential reasoning time complexity [115]. **Vocabulary modules:** NIF incorporates existing domain ontologies via vocabulary modules to provide best practices for NLP annotations for the whole breadth of the NLP domain, e.g. FISE (see below), ITS (Sect. 7.3.1), OLiA (Sect. 7.3.2), NERD [103].

7.3 Use Cases for NIF

7.3.1 Internationalization Tag Set 2.0

The *Internationalization Tag Set* (ITS) Version 2.0 is a W3C working draft, which is in the final phase of becoming a W3C recommendation. Among other things, ITS standardizes HTML and XML attributes which can be leveraged by the localization industry (especially language service providers) to annotate HTML and XML nodes with processing information for their data value chain. In the standard, ITS defines

[52] http://www.w3.org/TR/rdf-concepts/#section-Literals

[53] Available at http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/version-1.0/.

[54] See e.g. http://clarkparsia.com/pellet/icv/.

19 *data categories*[55], which provide a shared conceptualization by the W3C working group and its community of stakeholders. An example of three attributes in an HTML document is given here:

```
<html><body><h2 translate="yes">Welcome to <span
   its-ta-ident-ref="http://dbpedia.org/resource/Dublin" its-within-text="yes"
   translate="no">Dublin</span> in
   <b translate="no" its-within-text="yes">Ireland</b>!</h2></body></html>
```

As an outreach activity, the working group evaluated *RDFa*[56] to create a bridge to the RDF world, but concluded that the format was not suitable to serve as a best practice for RDF conversion. The main problem was that the defined ITS attributes annotate the text within the HTML nodes, but RDFa only has the capability to annotate resources with the text in the node as an object. RDFa lacks subject URIs, which refer to the text within the tags. Although it is theoretically possible to extract provenance information (i.e. offsets and position in the text), the RDFa standard does not include this use case and current RDFa parsers (with the exception of *viejs.org*) do not implement such an extraction.

In a joint effort, the ITS 2.0 RDF ontology[57] was developed using NIF, which was included within the proposed standard alongside an algorithm for a round-trip conversion of ITS attributes to NIF[58] (simple granularity). Provenance can be kept with an XPointer/XPath fragment identifier.

```
@base <http://example.com/nif.ttl#> .
<char=0,29> a nif:Context , nif:RFC5147String ;
    nif:beginIndex "0"  ;
    nif:endIndex   "29" ;
    nif:isString "Welcome␣to␣Dublin␣in␣Ireland!" .

<char=11,17> a  nif:RFC5147String ;
    nif:beginIndex "11" ;
    nif:endIndex   "17" ;
    nif:anchorOf         "Dublin" ;
    itsrdf:translate     "no";
    itsrdf:taIdentRef    dbpedia:Dublin ;
    # needed provenance for round-tripping
    prov:wasDerivedFrom  <xpath(/html/body[1]/h2[1]/span[1]/text()[1])> ;
    nif:referenceContext <char=0,29> .
```

NIF successfully creates a bridge between ITS and RDF and a round-trip conversion was recently implemented as a proof-of-concept. Therefore, NIF can be expected to receive a wide adoption by machine translation and industrial language service providers. Additionally, the ITS Ontology provides well modeled and accepted properties, which can in turn be used to provide best practices for NLP annotations.

[55] http://www.w3.org/TR/its20/#datacategory-description

[56] http://www.w3.org/TR/rdfa-syntax/

[57] http://www.w3.org/2005/11/its/rdf#

[58] ttp://www.w3.org/TR/its20/#conversion-to-nif

7.3.2 OLiA

The *Ontologies of Linguistic Annotation* (OLiA) [25][59] provide stable identifiers for morpho-syntactical annotation tag sets, so that NLP applications can use these identifiers as an interface for interoperability. OLiA provides *Annotation Models (AMs)* for fine-grained identifiers of NLP tag sets, such as *Penn*[60]. The individuals of these annotation models are then linked via `rdf:type` to coarse-grained classes from a *Reference Model (RM)*, which provides the interface for applications. The coverage is immense: OLiA comprises over 110 OWL ontologies for over 34 tag sets in 69 different languages, the latest addition being the Korean *Sejong tagset*. The benefit for application developers is three-fold:

1. **Documentation.** OLiA allows tagging with URIs (e.g. http://purl.org/olia/penn.owl#DT) instead of just short cryptic strings such as ”DT”. Developers who are unfamiliar can open the URL in an ontology browser and read the included documentation collected from the literature.
2. **Flexible Granularity.** For a wide range of NLP tools who built upon POS tags, very coarse-grained tags are sufficient. For example for keyword extraction, entity recognition and lemmatization, it is often not necessary to distinguish between singular/plural or common/proper noun. OLiA maps all four tags to a common class `olia:Noun`. Such a mapping exists for almost all tags and can be easily reused by developers for a wide range of tag sets.
3. **Language Independence.** AMs for different languages are mapped to the common RM providing an abstraction across languages.

NIF provides two properties: `nif:oliaLink` links a `nif:String` to an OLiA-AM. Although a reasoner could automatically deduce the abstract type of each OLiA individual from the RM, it was a requirement that the coarse-grained types should be linked redundantly to the strings as well in case reasoning services are not available or would cause high overhead. Therefore, an OWL annotation property `nif:oliaCategory` was created as illustrated in the following example.

```
<char=342,345> a nif:String, nif:RFC5147String ;
    nif:oliaLink        penn:NNP  ;
    nif:oliaCategory    olia:Noun , olia:ProperNoun .
# deducable by a reasoner:
penn:NNP       a olia:Noun,  olia:ProperNoun .
```

The NLP2RDF project provides conversions of the OLiA OWL files to CSV and Java HashMaps for easier consumption.[61] Consequently, queries such as ‘Return all strings that are annotated (i.e. typed) as `olia:PersonalPronoun` are possible, regardless of the underlying language or tag set.

All the ontologies are available under an open license.[62]

[59] http://purl.org/olia

[60] http://purl.org/olia/penn.owl

[61] `http://olia.nlp2rdf.org/owl/{Penn.java|penn.owl.csv|penn-link.rdf.csv}`

[62] http://sourceforge.net/projects/olia/

	NS	**NSI**	**OA**	**UC**
# triples	477 250 589	316 311 355	577 488 725	607 563 176
# generated URIs	76 850 241	42 880 316	169 849 625	189 342 046
# percentage	100%	66.28%	121.00%	127.30%
# percentage URIs	100%	55.79%	221.01%	246.38%

Table 1 Comparison of triple count and minted URIs. Percentage relative to NS. (NS = NIF Simple, NSI = NIF Simple Ideal, OA = Open Annotation, UC = UIMA Clerezza).

7.4 Qualitative Comparison with other Frameworks and Formats

In [55, 56], the *Graph Annotation Framework (GrAF)* was used to bridge the models of UIMA and GATE. GrAF is the XML serialization of the ISO standard *Linguistic Annotation Framework* (LAF) [67]. GrAF is meant to serve as a pivot format for conversion of different annotation formats and is able to allow a structural mapping between annotation structures. LAF/GrAF is very similar to the Open Annotation effort.

Extremely Annotational RDF Markup (EARMARK, [91]) is a stand-off format to annotate text with markup (XML, XHTML) and represent the markup in RDF including overlapping annotations. The main method to address content is via ranges that are similar to the NIF URI scheme. *TELIX* [107] extends SKOS-XL[63] and suggests RDFa as annotation format. We were unable to investigate *TELIX* in detail, because neither an implementation nor proper documentation was provided. In Section 7.3.1, we have argued already that RDFa is not a suitable format for NLP annotations in general. The usage of SKOS-XL by TELIX only covers a very small part of NLP annotations, i.e. lexical entities.

With the early *Tipster* and the more modern *UIMA* [32], *GATE* [27], *Ellogon*, *Heart-of-Gold* and *OpenNLP*[64] a number of comprehensive NLP frameworks already exist. NIF, however, focuses on interchange, interoperability as well as decentralization and is complementary to existing frameworks. Ultimately, NIF rather aims at establishing an ecosystem of interoperable NLP tools and services (including the ones mentioned above) instead of creating yet another monolithic (Java-)framework. By being directly based on RDF, Linked Data and ontologies, NIF also includes crucial features such as *annotation type inheritance* and *alternative annotations*, which are cumbersome to implement or not available in other NLP frameworks [108]. With its focus on conceptual and access interoperability NIF also facilitates *language resource* and *access structure* interchangeability, which is hard to realize with existing frameworks. NIF does not aim at replacing NLP frameworks, which are tailored for high-performance throughput of terabytes of text; it rather aims to ease access to the growing availability of heterogeneous NLP web services as, for example, already provided by *Zemanta* and *Open Calais*.

[63] http://www.w3.org/TR/skos-reference/skos-xl.html

[64] http://opennlp.apache.org

7.5 Lessons Learned, Conclusions and Future Work

Our evaluation of NIF since the publication of NIF 1.0 in the developers study has been accompanied by extensive feedback from the individual developers and it was possible to increase ontological coverage of NLP annotations in version 2.0, especially with the ITS 2.0 / RDF Ontology, NERD [103], FISE and many more ontologies that were available. Topics that dominated discussions were scalability, reusability, open licenses and persistence of identifiers. Consensus among developers was that RDF can hardly be used efficiently for NLP in the internal structure of a framework, but is valuable for exchange and integration. The implementation by Apache Stanbol offered a promising perspective on this issue as they increased scalability by transforming the identifiers used in OLiA into efficient Java code structures (enums). Hard-compiling ontological identifiers into the type systems of Gate and UIMA seems like a promising endeavour to unite the Semantic Web benefits with the scalability requirements of NLP. A major problem in the area remains the URI persistence. Since 2011 almost all of the mentioned ontologies either changed their namespace and hosting (OLiA and NIF itself) or might still need to change (Lemon, FISE), which renders most of the hard-coded implementations useless.

8 Summary and Recommendations

It is often said that the nice thing about standards is that there are so many of them, and this is certainly true for standards related to linguistic annotation. In addition to standards that have appeared over the past 30 years, numerous other formats have been developed and used by multiple projects, occasionally becoming accepted as *de facto* standards (e.g., the Penn Treebank bracketed format for syntax and its part-of-speech labels for English, CoNLL IOB for dependency analysis) for representing one linguistic phenomenon or another. To this day, anyone undertaking an annotation project can choose from multiple standards for representing the information added to a language resource, and no single option is necessarily superior to the others. However, the work that has been done on standards for linguistically annotated resources, although it has not led to a single, definitive solution that fits every case, has led to understanding of some best practices that can guide choices, especially for the developer of a new annotation scheme.

Perhaps the greatest lesson that 30 years of standards development has taught us is to separate the choices related to *annotation content* (i.e., linguistic categories and the names that will be used for them) from the choice of *representation format*. There exist several good representation formats, the most recently developed of which typically use the standoff approach and are manifestations of a graph-based data model (e.g., GrAF, NIF, RDF and any of its variant representations) and therefore trivially mappable, and converters among most of these formats are increasingly available.[65] Of course, the choice of representation format has to be made in the context of the software that will be used with the annotations (both creating and using them) . It also requires a decision among inline, standoff, and hybrid standoff annotation (see Part I, Chapter III for a discussion). Inline annotation is usually the least satisfactory approach for the reasons outlined in Chapter III, but it is also the easiest to process, either using available software such as XML parsers or writing relatively simple programs. A hybrid standoff approach typically relies on a fixed tokenization that is represented with XML elements, the former its drawback and the latter its appeal due to ease of referencing. Standoff allows for the most flexibility, but demands special processing to refer to and access data via offsets (at least as long as string references in formats like XML are problematic). Formats such as CoNLL IOB are very easy to process but pose problems for hierarchical annotations and, like hybrid methods, rely on a fixed tokenization. However, CoNLL-U (see Section 5 addresses some of these problems, including variant tokenizations, and in general demonstrates the degree to which standards for language resources are converging on common practices.

At the same time, standards for annotation content are far less well developed, and there are few, if any, widely accepted solutions. However, it is clear that the eventual solution will involve web-based repositories to which annotations can refer, in order to achieve uniformity among the concepts (if not the labels) that are applied. A great deal of work remains to be done to develop adequate coverage for the full range of linguistic phenomena within such repositories, as well as to find systematic ways to avoid reinvention and uncontrolled extension and, ultimately, link the various repositories to refer to an accepted set of common concept. At this point, awareness of what exists and the development process the community is pursuing, and utilizing repositories and inventories to the extent possible, is the best one can do.

The goal of this chapter has been to provide an overview of the state-of-the-art in standards development for language resources, examining issues in linguistic annotation of both text as well as spoken data. An attempt has been made to provide the context within which these annotation schemes were developed, along with an understanding of the considerations that have driven standards development and the current state of the standards landscape. Despite the lack of a single, generally applicable standard for representing linguistic annotations at this time, the situation is dramatically improved over what it was even 20 years ago, and convergence of practice is apparent. Improved and/or refined solutions are likely to emerge relatively rapidly over the next 10-20 years, hopefully enabling a huge increase in the availability, usability, and inter-connectedness of linguistically annotated resources.

[65] Note that converters from many graph-based formats to CoNLL IOB exist, but the reverse conversion from CoNLL IOB into these formats is significantly more challenging.

References

1. Allen, J., Core, M.: DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1). Technical Report. University of Rochester, Rochester, NY (1997). URL http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/
2. Allwood, J.: On dialogue cohesion. Gothenburg Papers in Theoretical Linguistics 65 (1992). Gothenburg University, Department of Linguistics
3. Auer, S., Hellmann, S.: The web of data: Decentralized, collaborative, interlinked and interoperable. In: LREC (2012)
4. Austin, P.K., Grenoble, L.A.: Current trends in language documentation. Language Documentation and Description **4** (2007)
5. Bigi, B., Hirst, D.: SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In: Speech Prosody, pp. 1–4. Shanghai, China (2012). URL https://hal.archives-ouvertes.fr/hal-00983699
6. Bird, S., Klein, E.: Phonological events. Journal of Linguistics **26**, 33–56 (1990)
7. Bird, S., Liberman, M.: A formal framework for linguistic annotation. Speech Communication **33**(1-2), 23–60 (2001)
8. Boersma, P., Weenink, D.: Praat, a system for doing phonetics by computer. Glot International **5**(9/10), 341–345 (2001)
9. Breen, M., Dilley, L.C., Kraemer, J., Gibson, E.: Inter-transcriber agreement for two systems of prosodic annotation: Tobi (tones and break indices) and rap (rhythm and pitch). Corpus Linguistics and Linguistic Theory **8**(2), 277–312 (2012)
10. Broeder, D., Schuurman, I., Windhouwer, M.: Experiences with the isocat data category registry. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 4565–4568. European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
11. Buchholz, S., Marsi, E.: CoNLL-X shared task on multilingual dependency parsing. In: Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL), pp. 149–164 (2006)
12. Bunt, H.: Context and dialogue control. Think Quarterly 3 (1) pp. 19–31 (1994)
13. Bunt, H.: Dialogue pragmatics and context specification. In: H. Bunt, W. Black (eds.) Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatics., pp. 81–150. John Benjamins, Amsterdam (2000)
14. Bunt, H.: A methodology for designing semantic annotation languages exploring semantic-syntactic iso-morphisms. In: A. Fang, N. Ide, J. Webster (eds.) Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010), pp. 29–46. Department of Chinese, Translation and Linguistics, City Univesity of Hong Kong, Hong Kong (2010)
15. Bunt, H.: Introducing abstract syntaxt + semantics in semantic annotation, and its consequences for the annotation of time and events. In: E. Lee, A. Yoon (eds.) Recent Trends in Language and Knowledge Processing, pp. 157–204. Hankookmunhwasa, Seoul (2011)
16. Bunt, H., Alexandersson, J., Carletta, J., Choe, J.W., Fang, A.C., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., Traum, D.: Towards an iso standard for dialogue act annotation. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (2010)
17. Bunt, H., Alexandersson, J., Choe, J.W., Fang, A.C., Hasida, K., Petukhova, V., Popescu-Belis, A., Traum, D.: Iso 24617-2: A semantically-based standard for dialogue annotation. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (2012)
18. Bunt, H., Palmer, M.: Conceptual and representational choices in defining an iso standard for semantic role annotation. In: H. Bunt (ed.) Proceedings of the 9th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9), pp. 41–50. Association for Computational Linguistics, Potsdam, Germany (2013). URL http://www.aclweb.org/anthology/W13-0500

19. Bunt, H., Palmer, M.: Conceptual and representational choices in defining an iso standard for semantic role annotation. In: H. Bunt (ed.) Proceedings of the 9th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9), pp. 41–50. Association for Computational Linguistics, Potsdam, Germany (2013). URL http://www.aclweb.org/anthology/W13-0500
20. Bunt, H., Pustejvosky, J.: Annotating event and temporal quantification. In: Proceedings of the Fifth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation ISA-5, pp. 15–22 (2010)
21. Buschmeier, H., Wlodarczak, M.: Textgridtools: A textgrid processing and analysis toolkit for python. In: Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013), pp. 152–157 (2013)
22. Carletta, J., Dahlbäck, N., Reithinger, N., Walker, M.A.: Standards for dialogue coding in natural language processing. Tech. Rep. Report no. 167 (1997). Report from Dagstuhl seminar number 9706
23. Carletta, J., Isard, S., Kowtko, J., Doherty-Sneddon, G.: HCRC dialogue structure coding manual (1996). Technical Report HCRC/TR-82
24. Corpus Encoding Standard. http://www.cs.vassar.edu/CES/CES1.html (1994). URL http://www.cs.vassar.edu/CES/CES1.html
25. Chiarcos, C.: Ontologies of linguistic annotation: Survey and perspectives. In: LREC. European Language Resources Association (2012)
26. Cinková, S.: From propbank to engvallex: Adapting the propbank-lexicon to the valency theory of the functional generative description. In: Proceedings of the 6th Edition of International Conference on Language Resources and Evaluation (LREC 2006), pp. 2170–2175 (2006)
27. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: ACL (2002). DOI 10.3115/1073083.1073112. URL http://www.aclweb.org/anthology/P02-1022
28. Dhillon, R., Bhagat, S., Carvey, H., Schriberg, E.: Meeting recorder project: dialogue labelling guide. (2004). ICSI Technical Report TR-04-002.
29. Eckle-Kohler, J., Gurevych, I., Hartmann, S., Matuschek, M., Meyer, C.M.: Uby-lmf - exploring the boundaries of language-independent lexicon models. In: G. Francopoulo (ed.) LMF Lexical Markup Framework, chap. 10, pp. 145–156. ISTE - HERMES - Wiley, London, UK (2013)
30. Farrar, S., Langendoen, D.T.: A Linguistic Ontology for the Semantic Web. GLOT International **7**, 97–100 (2003)
31. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA (1998)
32. 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)
33. Fillmore, C., Baker, C., Sato, H.: Framenet as a ”net”. In: Proceedings of the 4th Edition of International Conference on Language Resources and Evaluation (LREC 2004), pp. 1091–1094 (2004)
34. Fillmore, C.J.: The case for case. In: E. Bach, R. Harms (eds.) Universals in Linguistic Theory, pp. 1–89. Holt, Rinehart, and Winston (1968)
35. Francopoulo, G. (ed.): LMF: Lexical Markup Framework. London: Wiley-ISTE (2013)
36. Gibbon, D.: Time types and time trees: prosodic mining and alignment of temporally annotated data. In: S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, J. Schlieer (eds.) Methods in Empirical Prosody Research, pp. 281–209. Walter de Gruyter, Berlin (2006)
37. Gibbon, D.: Modelling gesture as speech: A linguistic approach. Pozna? Studies in Contemporary Linguistics **47**, 470–508 (2011)
38. Gibbon, D., Mertins, I., Moore, R.: Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation. The Springer International Series in Engineering and Computer Science. Springer US (2000). URL http://books.google.com/books?id=Ntb0T7gfIn8C

39. Gibbon, D., Moore, R., Winski, R. (eds.): Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter (1997)
40. G?owi?ska, K., Przepirkowski, A.: The design of syntactic annotation levels in the national corpus of polish. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp. 19–21. European Language Resources Association (ELRA), Valletta, Malta (gg)
41. Grishman, R.: TIPSTER Architecture Design Document Version 2.2. Tech. rep., Defense Advanced Research Projects Agency (1996)
42. Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C.M., Wirth, C.: Uby - A Large-Scale Unified Lexical-Semantic Resource. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), pp. 580–590. Avignon, France (2012)
43. Hellmann, S., Lehmann, J., Auer, S.: Linked-data aware uri schemes for referencing text fragments. In: EKAW 2012, LNCS 7603. Springer (2012)
44. Hirst, D., Di Cristo, A.: Intonation Systems: A Survey of Twenty Languages. Cambridge University Press (1998). URL http://www.google.com.sg/books?id=LClvNiI4k0sC
45. Ide, N., Baker, C., Fellbaum, C., Passonneau, R.: The manually annotated sub-corpus: A community resource for and by the people. In: Proceedings of the ACL 2010 Conference Short Papers, pp. 68–73. Association for Computational Linguistics, Uppsala, Sweden (2010)
46. Ide, N., Bonhomme, P., Romary, L.: XCES: An XML-based encoding standard for linguistic corpora. In: Proceedings of the Second International Language Resources and Evaluation Conference (LREC'00) (2000)
47. Ide, N., Pustejovsky, J.: What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability. In: Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong (2010)
48. Ide, N., Pustejovsky, J., Suderman, K., Verhagen, M.: The Language Application Grid Web Service Exchange Vocabulary. In: Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT). Dublin (2014)
49. Ide, N., Romary, L.: Standards for language resources. In: Proceedings of the IRCS Workshop on Linguistic Databases, pp. 141–9. Philapdelphia, Pa. (2001)
50. Ide, N., Romary, L.: Outline of the International Standard Linguistic Annotation Framework. In: Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right, pp. 1–5 (2003)
51. Ide, N., Romary, L.: International standard for a linguistic annotation framework. Natural Language Engineering **10**(3-4), 211–225 (2004)
52. Ide, N., Romary, L.: International standard for a linguistic annotation framework. Natural Language Engineering pp. 10, 211225 (2004)
53. Ide, N., Romary, L.: A registry of standard data categories for linguistic annotation. In: In Proceedings of the Fourth Language Resources and Evaluation Conference (LREC), pp. 135–139. Lisbon (2004)
54. Ide, N., Romary, L.: Towards International Standards for Language Resources. In: L. Dybkjaer, H. Hemsen, W. Minker (eds.) Evaluation of Text and Speech Systems, pp. 263–84. Springer (2007)
55. Ide, N., Suderman, K.: GrAF: A Graph-based Format for Linguistic Annotations. In: Proceedings of the Linguistic Annotation Workshop (LAW), pp. 1–8. Association for Computational Linguistics (2007)
56. Ide, N., Suderman, K.: The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging. Language Resources and Evaluation **48**(3), 395–418 (2014)
57. Ide, N., Veronis, J.: Multext: Multilingual text tools and corpora. In: COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics (1994). URL http://aclweb.org/anthology/C94-1097
58. Ide, N., Veronis, J.: Encoding dictionaries. In: N. Ide, J. Veronis (eds.) The Text Encoding Initiative: Background and Context. Kluwer Academic Publishers, Dordrecht (1995)

59. International Organization for Standardization: ISO 8879:1986: Information processing — Text and office systems — Standard Generalized Markup Language (SGML). ISO, Geneva (1986)
60. ISO: ISO 8601:2004 Data elements and interchange formats – Information interchange – Representation of dates and times. ISO, Geneva (2004)
61. ISO: ISO 24612:2012 Language resource management – Linguistic annotation framework (LAF). ISO, Geneva (2012). ISO Working Group:TC 37/SC 4/WG 1, Convenor and Project leader: Nancy Ide
62. ISO: ISO 24617-1:2012 Language resource management - Semantic annotation framework - Part 1: Time and events (SemAF-Time, ISO-TimeML). ISO, Geneva (2012). ISO Working Group:TC 37/SC 4/WG 2, Editors: James Pustejvosky (chair), Harry Bunt, Kiyong Lee (convenor and project leader), Bran Boguraev, and Nancy Ide in cooperation with the TimeML Working Group, http://www.timeml.org.
63. ISO: ISO 24617-2:2012 Language resource management - Semantic annotation framework - Part 2: Dialogue acts (SemAF-DA). ISO, Geneva (2012). ISO Working Group:TC 37/SC 4/WG 2 Convenor: Kiyong Lee, Project leader: Harry Bunt
64. ISO: Language Resource Management - Linguistic Annotation Framework. iso 24612 (2012)
65. ISO: ISO 24617-4:2014 Language resource management - Semantic annotation framework - Part 4: Semantic roles (SemAF-SR). ISO, Geneva (2014). ISO Working Group:TC 37/SC 4/WG 2 Convenor: Kiyong Lee, Project leader: Martha Palmer, Writers: Martha Palmer (USA), Collin Baker (USA), Claire Bonial (USA), Harry Bunt (Holland), Katrin Erk (USA, Germany),Olga Petukhova (Germany), James Pustejovsky (USA), Zdenka Uresova (the Czech Republic), Nianwen Xue (USA, China)
66. ISO: ISO 24617-7:2014 Language resource management - Part 7: Spatial information (ISOspace). ISO, Geneva (2014). ISO Working Group: TC 37/SC 4/WG 2, Project leaders: James Pustejovsky and Kiyong Lee, supported by the ISOspace Working Group headed by James Pustejvosky at Brandeis University, Waltham, MA, U.S.A. The following is the homepage for the ISO-Space project <https://sites.google.com/site/wikiisospace/>.
67. ISO 24612:2012: Language resource management, Linguistic annotation framework (LAF). ISO, Geneva, Switzerland (2012)
68. Katz, G.: Annotating temporal and event quantification. Annotating, Extracting and Reasoning about Time and Events pp. 88–106 (2007)
69. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: A Large-scale Classification of English Verbs. Language Resources and Evaluation **42**, 21–40 (2008)
70. Kipper-Schuler, K.: Verbnet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania (2005)
71. Klessa, K., Gibbon, D.: Annotation pro + tga: Automation of speech timing analysis. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
72. Knuth, D.E.: Literate Programming. CSLI Lecture Notes Number 27 (1992)
73. Kübler, S., McDonald, R., Nivre, J.: Dependency Parsing. Morgan and Claypool (2009)
74. Laurent, R.: Tei and lmf crosswalks. digital humanities (forthcoming)
75. Lee, K.: Formal semantics for temporal annotation. Lecture Notes for CIL18 (2008)
76. Lee, K.: A cmpositional interval semantics for temporal annotation. In: E. Lee, A. Yoon (eds.) Recent Trends in Language and Knowledge Processing, pp. 157–204. Hankookmunhwasa, Seoul (2011). Presented at the workshop on language and knowledge processing, Pusan National University, in summer 2008.
77. Lee, K.: The annotation of measure expressions in ISO standards. In: H. Bunt (ed.) Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11). QMUL, London (2015). A satellite workshop of IWCS 2015, London, U.K.
78. Lee, K., Romary, L.: Towards interoperability of ISO standards for language resource management. In: A.C. Fang, N. Ide, J. Webster (eds.) Proceedings of Language Resources and Interoperability, The Second International Conference on Global Interoperability for Language Resources (ICGL201), pp. 95–104. Hong Kong (2010)

79. Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., Wellner, B.: Spatialml: Annotation scheme, corpora, and tools. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (2008). Http://www.lrec-conf.org/proceedings/lrec2008/
80. de Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.D.: Universal Stanford Dependencies: A cross-linguistic typology. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), pp. 4585–4592 (2014)
81. de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC) (2006)
82. de Marneffe, M.C., Manning, C.D.: The Stanford typed dependencies representation. In: Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation, pp. 1–8 (2008)
83. McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., Lee, J.: Universal dependency annotation for multilingual parsing. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 92–97 (2013)
84. McDonald, R., Petrov, S., Hall, K.: Multi-source transfer of delexicalized dependency parsers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 62–72 (2011)
85. Mcneill, D. (ed.): Language and Gesture: Window into Thought and Action. Cambridge University Press (2000)
86. Mehler, A., Romary, L., Gibbon, D. (eds.): Handbook of Technical Communication. Handbooks of Applied Linguistics. De Gruyter Mouton, Berlin and Boston (2012)
87. MITRE: SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language. The MITRE Corporation (2009). Version 3.1, October 1, 2009, Contact: cdoran@mitre.org.
88. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007, pp. 915–932 (2007)
89. Nivre, J., Hall, J., Nilsson, J.: Maltparser: A data-driven parser-generator for dependency parsing. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 2216–2219 (2006)
90. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles **31(1)**, 71–0106 (2005)
91. Peroni, S., Vitali, F.: Annotations with earmark for arbitrary, overlapping and out-of order markup. In: U.M. Borghoff, B. Chidlovskii (eds.) ACM Symposium on Document Engineering, pp. 171–180. ACM (2009)
92. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC) (2012)
93. Petukhova, V., Bunt, H.: The independence of dimensions in multidimensional dialogue act annotation. In: Proceedings NAACL HLT Conference, Boulder, Colorado (2009)
94. Petukhova, V., Bunt, H., Schiffrin, A.: Lirics semantic role annotation: Design and evaluation of a set of data categories. In: Proceedings of the 6th Edition of International Conference on Language Resources and Evaluation (LREC 2008). Marrakech (2007)
95. Petukhova, V., Prévot, L., Bunt, H.: Discourse relations in dialogue. In: Proceedings 6th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-6). Oxford, UK (2011)
96. Popescu-Belis, A.: Dialogue Acts: One or More Dimensions? ISSCO Working Paper 62. ISSCO, Geneva (2005). URL http://www.issco.unige.ch/publicaitons/working-papers/papers/apb-issco-wp62b.pdf
97. Pratt-Hartmann, I.: From TimeML to interval temporal logic. In: H. Bunt (ed.) Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7), pp. 166–180. Tilburg, The Netherlands (2007)

98. Przepiórkowski, A.: TEI P5 as an XML Standard for Treebank Encoding, pp. 149–160 (2009)
99. Pustejovsky, J., Ingria, R., Saurí, R., Castaño, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., Mani, I.: The specification language TimeML. In: I. Mani, J. Pustejvosky, R. Gaizauskas (eds.) The Language of Time: a Reader, pp. 545–557. Oxford University Press, Cambridge (2005)
100. Pustejovsky, J., Lee, K., Bunt, H., Romary, L.: Iso-timeml: An international standard for semantic annotation. In: Proceedings of LREC2010. Malta (2010)
101. Pustejovsky, J., Lee, K., Bunt, H., Romary, L.: ISO-TimeML: an international standard for semantic annotation. In: Proceedings of LREC 2010. La Valette, Malta (2010)
102. Pustejvosky, J., Gaizauskas, R., Saurí, R., Setzer, A., Ingrai, R.: Annotation guideline to TimeML 1.0 (2002). Available at `<http://timeml.org>`
103. Rizzo, G., Troncy, R., Hellmann, S., Bruemmer, M.: NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In: LDOW (2012)
104. Romary, L.: TBX goes TEI - implementing a TBX basic extension for the text encoding initiative guidelines. CoRR **abs/1403.0052** (2014)
105. Romary, L., Bonhomme, P.: Parallel alignment of structured documents. In: Parallel Text Processing, pp. 201–217. Springer (2000)
106. Rossini, N.: Reinterpreting Gesture as Language - Language in Action. IOS Press (2012)
107. Rubiera, E., Polo, L., Berrueta, D., Ghali, A.E.: Telix: An rdf-based model for linguistic annotation. In: ESWC (2012)
108. Schierle, M.: Language engineering for information extraction. Phd thesis, Universität Leipzig (2011)
109. Schiffrin, A., Bunt, H.: LIRICS Deliverable D4.3: Documented compilation of semantic data categories (2007). URL http://lirics.loria.fr
110. Schmidt, T.: A tei-based approach to standardising spoken language transcription. Journal of the Text Encoding Initiative (1) (2011)
111. Sperberg-McQueen, C., L. Burnard, L. (eds.): Guidelines for Electronic Text Encoding and Interchange. TEI P3. Text Encoding Initiative, Oxford, Providence, Charlottesville, Bergen (1994)
112. Szyma?ski, M., Bachan, J.: Interlabeller agreement on segmental and prosodic annotation of the jurisdict polish database. Speech and Language Technology **14/15**, 105–121 (2012)
113. TEI Consortium (ed.): Guidelines for Electronic Text Encoding and Interchange. TEI P5. Text Encoding Initiative, Oxford, Providence, Charlottesville, Bergen, Nancy (2003)
114. Teoh, A., Chin, S.: Transcribing the speech of children with cochlear implants: Clinical application of narrow phonetic transcriptions. American Journal of Speech and Language Pathology **18**(4), 388–401 (2009)
115. Tobies, S.: Complexity results and practical algorithms for logics in knowledge representation. Ph.D. thesis, TU Dresden (2001)
116. Tomaz, E., Fiser, D., Krek, S., Ledinek, N.: The jos linguistically tagged corpus of slovene. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Malta (2010)
117. Traum, D.: 20 questions on dialogue act taxonomies. Journal of Semantics 17(1) pp. 7–30. (2000)
118. Tsarfaty, R.: A unified morpho-syntactic scheme of Stanford dependencies. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 578–584 (2013)
119. Windhouwer, M.: RELcat: a Relation Registry for ISOcat data categories. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 3661 – 3664. Istanbul, Turkey (2012)
120. Windhouwer, M., Wright, S.E.: LMF and the Data Category Registry: principles and application. In: G. Francopoulo (ed.) LMF Lexical Markup Framework, chap. 10, pp. 41–50. ISTE - HERMES - Wiley, London, UK (2013)
121. Zeman, D.: Reusable tagset conversion using tagset drivers. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), pp. 213–218 (2008)

122. Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., Hajič, J.: HamleDT: To parse or not to parse? In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), pp. 2735–2741 (2012)