Multilingual Language Resources and Interoperability

Andreas Witt (andreas.witt@uni-tuebingen.de), Ulrich Heid (heid@ims.uni-stuttgart.de), Felix Sasaki (fsasaki@w3.org) and Gilles Sérasset (gilles.serasset@imag.fr)

Abstract. This article introduces the topic of "Multilingual Language Resources and Interoperability". We start with a taxonomy and parameters for classifying language resources. Later we provide examples and issues of interoperatability, and resource architectures to solve such issues. Finally we discuss aspects of linguistic formalisms and interoperability.

Keywords: language, resources, interoperability

1. Introduction

This special issue of *Language Resources and Evaluation*, entitled "Multilingual Language Resources and Interoperability", is composed of extended versions of selected papers from the COLING/ACL Workshop on *Multilingual Language Resources and Interoperability*, held in 2006, in Sydney (cf. Witt et al., 2006). This introduction does not attempt to provide a complete overview of this vast topic, but rather sketches the background against which the articles assembled in this issue are to be read. In particular, we examine the notions of (multilingual) language resources (section 2) and interoperability of resources (section 3), and assess resource architectures (section 4) and linguistic representation formalisms (section 5) with respect to their potential to support resource interoperability. This background provides a framework in which each paper in this issue of the journal is then situated.

2. Language Resources

Often, the term *language resources* is taken to refer to corpora and lexicons. This view is incomplete in several respects. Obviously, speech resources are not explicitly included in this definition. And even if we stick to non-speech resources (sometimes called "text resources" or "NLP resources"), as we intend to do it in this article, this simplistic view is still insufficient, because it excludes other vital aspects of the process of creating, representing, and maintaining language resources, such as the wide array of annotation tools (e.g., part-of-speech-taggers, morphological analysers, parsers, etc.) typically applied to archived

language data. Nor does it account for the wide variety of ways in which lexical knowledge can be structured: not only by lemmata or by graphic words, but also by concept (as in an ontology) or other properties such as pronunciation, valency, etc. Therefore, in this article we use a broader definition of the term that encompasses both *static* and *dynamic* resources, where static resources are inventories of data, and dynamic resources are tools that produce new data, for example linguistic annotations, corpus-based generated lexica, translations. Thus, corpora, lexicons, ontologies, terminology inventories etc. are regarded as static resources, while taggers, morphological analysers, parsers, tools for lexical data extraction, etc. are regarded as dynamic resources[1].

2.1. A simple taxonomy of language resources

Both static and dynamic resources may be *text-based* or *item-based*, depending on the size of the linguistic objects involved. For example, corpora are text-based static resources, whereas all kinds of lexicons and ontologies, which consist of collections of individual items, are item-based static resources. Similarly, taggers, morphology systems, word guessers, etc. are all item-based dynamic resources because they produce linguistic information associated with single items; while parsers, natural language understanding tools, and most machine-learning-based tools are among the dynamic text-based resources. Obviously, there are resources that are a mixture of item-based and text-based resources, such as tools that manipulate both lexicon and corpus.

2.2. Further parameters for classifying language resources

Along with this simple taxonomy of language resources, there are other aspects which play a role in resource definition and subclassification. Most importantly, resources can provide more or less linguistic interpretation. For example *raw data*, such as unannotated text, contains no interpretive information apart from the fact that it has been selected for a specific purpose. This applies equally to word lists or tools that produce these lists, as well as handwritten texts, output from speech recognition, etc. What we call '*primary data*', in the following discussion, is interpretative in the sense that, for example, a person

[1] With resources being distributed, for example, over the web, sometimes this distinction is difficult to make, for example, online lexicons generated and updated automatically each time they are accessed. Nevertheless, the distinction holds if we consider their purpose and role for processing linguistic resources.

transcribing speech has decided to have heard a given word or word sequence, and not another one.

Finally, there is a vast array of interpretations that can be added to primary data by applying annotation tools and by data enhancement (e.g. adding items to a lexicon)[2]. Annotation, here, encompasses the manual or automatic assignment of interpretative data to raw or primary data, and it covers both metadata annotation (e.g. data about the source of items and of their linguistic description, about the way in which the material was collected, etc.) and linguistic annotation. The latter is obviously available for many different levels of linguistic description (morphology, syntax, lexical semantics, etc.), and annotations from these levels may exist individually or in combination. Figure 1 summarizes the above discussion graphically, combining our simple taxonomy from section 2.1 with the criterion of how much interpretation is provided in a resource. The lower part of the figure concerning specifications will be discussed in section 3.2.

Annotated corpora, lexicons which are more than mere word lists, and tools to produce such static resources may differ in ways other than the levels of linguistic description under study. For example, each level of linguistic description may represent different theoretical or methodological approaches and resources may focus on linguistic objects of very different granularities: morphemes, word forms, chunks, phrases, sentences, texts, etc. This leads to the possibility of considerable variation in the ways in which different resources describe linguistic objects, but such variation may also exist within a single resource, for example, a corpus with concurrent annotations from one level of description, or a dictionary that contains descriptions according to different approaches or for different applications.

2.3. Multilingual resources

Multilingual resources typically also contain multiple annotations: they describe linguistic objects from different languages, along with individual interpretative annotations and, possibly, annotations creating relations between languages that add a contrastive interpretation[3]. This general description fits not only parallel corpora and bilingual

[2] In corpus linguistics, there are proponents of an analysis of primary data without annotations of any kind, using the argument that tools or interactive procedures that add annotations into a text may introduce interpretations which the analyst would not want to share. This view is to some extent related to (or at least quite frequently shared by) the "corpus-driven" paradigm of corpus linguistics (cf. Tognini-Bonelli, 2001).

[3] (Resnik et al., 1999) use the Bible as a multilingual parallel corpus; it is the text available in the largest number of different languages. Other aspects of multilingual

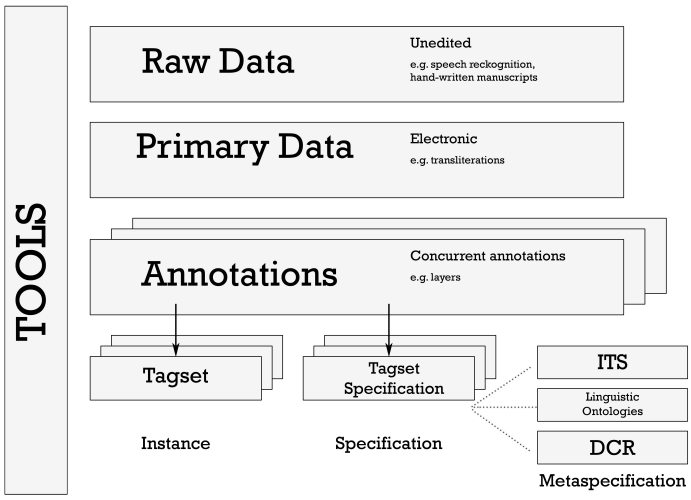

Figure 1. A simple classification of linguistic resources

dictionaries, but also collections of texts from different languages and alignment and annotation transfer tools. Obviously, contrastive links in dictionaries tend to be more explicitly classified than, for example, links in sentence or word alignment. The least precise contrastive link is found in non-parallel multilingual collections, such as, for example, comparable corpora.

Bilingual and multilingual dictionaries pose additional problems with respect to the descriptive devices used: it is necessary to provide coherent descriptions of the individual languages (i.e. the descriptive system of each language section must be coherent), and, in addition, it is necessary to ensure that parallel, or comparable, or at least systematically relatable classifications are used across languages. Such problems arise, for example, in work on collocations, or in multilingual terminology work (cf. Lyding et al., 2006). Similar issues are addressed, for example, in the PARGRAM project[4] where parallel grammars for several typologically different languages, as well as the respective lexical resources

text collections and their use in NLP are addressed in the contribution by Seretan and Wehrli (in this issue).

[4] PARGRAM's URL is, as of December 2008: `http://www2.parc.com/isl/groups/nltt/pargram/`

are created. Their parallel is based on the use of *Lexical Functional Grammar* (Kaplan and Bresnan, 1982), and on the explicit interest in coordinating the definition and use of descriptive devices across the languages.

In recent years the Web has been used extensively for creating large, multilingual corpora, used for example as training material for machine translation. For such applications, information about the language of a Web page or its character encoding is crucial. Current approaches for harvesting Web corpora usually do not make use of declarative information in this area, which would be available e.g. from the protocol (e.g. an HTTP-header) or the content itself (e.g. an encoding or language declaration). See (Emerson and O'Neil, 2006) for a description of these approaches. Nevertheless the amount of Web pages with correct information in this area is growing, and it can be expected that this will benefit the harvesting of multilingual Web corpora significantly. As a contribution for reaching this goal and the realization of the multilingual Web in general, the World Wide Web Consortium (W3C) is producing various outreach and training materials. See http://www.w3.org/International/ for further information.

2.4. Intermediate summary

Overall, the term *language resources* is rather broadly defined, even when speech resources are not included. It refers to entities that are multidimensional (static vs. dynamic, item-based vs. text-based, interpretative vs. non-interpretative) and includes numerous variation parameters (levels of linguistic description, "size" and type of linguistic objects treated, approach or theory followed, language(s) dealt with, etc.). Resource interoperability must deal with this complexity; however, it is obvious that not all of the aspects mentioned above can be addressed at the same time. Consequently, most of the discussion of interoperability of language resources in the following section focuses on particular aspects or specific combinations of aspects of the problem.

3. Interoperability of Language Resources

The most general definition of interoperability of language resources is the capability for these resources to interact or to work together. Interoperability may exist between static and dynamic resources as well as different static resources. This general notion of interoperability is used in computer science, to denote the capacity of programs, components, representations, data structures to interact.

3.1. Examples of resource interoperation

Typical examples of static and dynamic language resources interoperability involve the interaction between corpora and tools such as taggers, parsers, etc. Most such tools in turn must be interoperable with other static resources, e.g. lexicons.

Most interoperability among lexicons, corpora and tools is accomplished at the tool developer's side; as long as the same author(s) are responsible for all components, interoperability is simple to achieve. But often, tools are shared, corpora or lexicons are provided to other users and/or developed by (distributed) teams, or resources are used for other NLP applications than those for which they were originally intended. This is where interoperability can become problematic, potentially making it impossible to use the data and/or tools together as desired. Most of the remainder of this paper and the other papers in this issue address means to ensure that sharing and reuse of language resources is feasible and efficient.

Other examples of interoperability of static and dynamic resources concern the combination of corpora, lexicons and tools within one dimension of linguistic interpretation (e.g. the use of someone else's morphological analyser on my own tagged corpus) or across dimensions (e.g. the combination of a tagger and a chunker or parser, on a given corpus). Interoperability is also at stake when different static resources are combined, e.g. when different lexicons are merged, or when corpora from different sources are combined and common subsets created. Similarly, the interrelation between annotations from a corpus of one language with one of another language requires interoperability. For example, simple operations like counting words are hard to compare between languages if there is no clear definition of the concept word in the two languages and their relation.

There is a growing interest in resource sharing to achieve cost effectiveness: creating lexicons, annotating corpora or developing NLP tools are time-consuming, laborious and costly tasks. Thus, if NLP technology is to be used to a wider extent than today, it will be necessary to make resources available to a broad range of users, for example over the web. Current initiatives such as the European large-scale project CLARIN funded by the EC, and its national parallel projects, aim at exploring technical, but also and in particular conceptual foundations of scenarios for sharing and reusing language resources. (Váradi et al., 2008).

3.2. Issues of resource interoperability

The above-mentioned scenarios concern sharing and exchange of language resources in the broad sense or the combination of resources. These resources may be of different types and/or they may show the kinds of variation discussed in section 2. For example, they may be constructed according to different theories or approaches, or for different applications.

When resources are combined, descriptions from different levels must fit together; when annotations are interrelated, it must be ensured that the target text is annotated in the same way as the source text. In all cases, specifications of language resources (see the lower part of Figure 1) are the conceptual basis of interoperability: they should provide a formal description of the content of the resources and other aspects: items covered, descriptive dimensions (= attributes) appropriate for each type of item, appropriate values and their type or range for each descriptive dimension, relations between linguistic objects and/or annotations, etc. Examples of such specifications are tagset specifications and the pertaining annotators' guidelines, stylebooks and schemata of lexical resources, (meta)models for lexicons, and many more. Additionally, the choice of the resource architecture, the choice of a formalism on which it is based, as well as principles for the way in which resources are compared with respect to the criteria mentioned above, i.e. approaches to realizing interoperability, play a crucial role. In the following section, we will discuss several architectures for language resources and provide some basic principles for making interoperability between resources easier.

4. Resource Architectures and Interoperability

4.1. Examples of resource architectures

Interoperability is among the prime objectives of a paradigm of representation and processing of NLP data developed by Hermann Helbig (Helbig, 2001), called multilayered extended semantic networks (Multinet). This author starts his description of the representation system by discussing a number of general requirements for knowledge representation and knowledge processing. For him, interoperability, i.e. the possibility to combine knowledge representation and knowledge processing seamlessly, i.e. without specific interfaces, is one of these general requirements. In Helbig's list of desiderata there are two other requirements which are closely linked, namely homogeneity and communicability. Homogeneity of representations implies that one and the

same formalism is used to represent data from different levels of linguistic description, and, potentially, support inferences. Communicability targets documentation and thus the possibility to develop resources in a team, as well as allowing for the sharing of resources.

Helbig's model covers various aspects of NL understanding and, for example, also integrates a parser, a coreference resolution tool, as well as applications in information retrieval, information extraction and natural language database access.

Work on issues of interoperability has been carried out, among others, in the following areas:

- The design of integrated processing environments for NLP such as GATE (Cunningham, 2002), UIMA (Ferrucci and Lally, 2004), or Heart of Gold (Schäfer, 2006). These systems provide a platform for the combination of static and dynamic resources, typically in order to implement a processing pipeline or another modular software architecture that allows the user to derive linguistic representations at different levels by combining resources. The emphasis is on interface specifications.

- The design of multilayered annotation schemes for corpora, for example in the framework of the MATE and NITE projects (Carletta et al., 2003), which were European attempts to design a representation and query system for multiply annotated text and speech corpora in parallel. The *Annotation Graph* framework (Bird and Liberman, 2001) and Graph-based Format for Linguistic Annotations, GrAF (Ide and Suderman, 2007) were designed for similar purposes. Moreover the standardized XML-representation of feature structures, a joint TEI recommendation (Burnard and Bauman, 2007) and ISO standard (ISO 24610-1:2006, 2006), is applicable to represent multiply annotated (linguistic) data (Witt et al., 2009).[5] Multi-layered annotation schemes might be used, for example, to annotate the results of different analysers as layers of annotation (or, in case of alternatives at one level, as concurrent annotations) to a language resource. Thus a static resource can be designed which includes data from different levels of description and provides adequate means to compare and combine data from these levels. Here, the emphasis is on homogeneity, i.e. representing data from different levels by means of a common format (Wörner et al., 2006). This obviously also includes possibilities of jointly interrogating the different annotation layers.

[5] These approaches can also be used in combination, e.g. GrAF can use the TEI feature structures to represent annotation information.

– The design of architectures for lexical resources; typically, a complex NLP dictionary hosts data pertaining to different levels of linguistic description, on a per item basis, e.g. for individual words. Similar in spirit to multilayered corpus annotation (which aims at text-based static resources), several architectural proposals for lexical resources aim at item-based static resources covering different levels of description and, possibly concurrent classifications of lexical items. An example is Trippel's (2006) proposal for a graph-based lexicon model (Trippel, 2006), or the *Lexical Systems* proposal (Polguère, this issue). Both provide general frameworks which can be used to accommodate different, even diverging, lexical descriptions. Other models with a related objective are Papillon (Boitet et al., 2002) and the MILE proposal (Calzolari et al., 2002). The latter two are explicitly oriented towards multilinguality, i.e. they provide devices to express contrastive knowledge. MILE, in particular, is a clear example of how monolingual dictionaries are combined into multilingual ones: as in PARGRAM, the monolingual resources are constructed according to common principles.

– Work towards contents-wise (meta)standards for resource building, such as the Lexical Markup Framework (LMF, Francopoulo et al., 2006b, Francopoulo et al., 2006a), or, in the field of terminology, the Terminology Markup Framework (TMF[6]). These proposals specify in very general terms the basic building blocks of lexical or terminological resources and their interrelation, on the basis of the result of a consensus-based standardization process. Emphasis here is on the generality of the metamodels, which allows for different instantiations. For example, LMF makes no restrictions as to whether multilingual resources should be concept and interlingua-based, or whether they should be transfer-based, as is the case in some commercial symbolic Machine Translation (MT) systems. The (non-normative) LMF instantiation for bilingual and multilingual dictionaries caters for both approaches, and, additionally for translation memory data, i.e. for resources from example-based MT; it thus allows in principle for an exchange of data from all three approaches (cf. Soria et al. in this issue, for more details and an application).

– Work on identification of dynamic and static language resources. This is important in two applications: referencing linguistic resources and identifying dynamic and static resources for distributed

[6] TMF's URL is, as of December 2008: `http://www.loria.fr/projets/TMF/`

applications. For the former, a standardization proposal is currently under discussion within the International Standards Organization (ISO), as part of ISO TC37/SC4. tools that make use of part of speech tagging tools available online, which may in turn make use of lexical resources that are then generated and updated separately. Several questions arise concerning distributed processing, including 1) which information should be put into identifiers (e.g. the name of a resource, or a sub resources and query parameters) or other parts of a service request? This topic is discussed between the schools of RESTful and non-RESTful Web Services (Richardson and Ruby, 2007), not being specific to language resources, but to distributed information architectures in general. 2) Which protocol should be used, for example, HTTP or language resource community specific protocols? 3) How can we identify sub-parts of resources of different media, for example text, audio, or video? 4) How can we keep track of versioning of resources?

The above list follows our distinction between dynamic (GATE, UIMA, etc.) vs. static resources (e.g. corpora), and text-based resources (NITE etc.) vs. item-based ones (e.g. lexical models). A closer comparison of these different proposals from the point of view of how they actually deal with exchange and combination reveals that they can be classified as following one of two major philosophies, which we call the "transfer" and "interlingua" philosophies, deliberately using terminology from the machine translation field.

4.2. Two philosophies for resource interoperability

To exemplify the two philosophies for interoperability, we assume a situation where two lexical resources are to be merged, or where corpus data with different annotations from the same level of linguistic description have to be merged. The same scenario might also arise when applying a tool from one site to a corpus from another.

In this situation, a *transfer philosophy of interoperability* will analyze both representations at hand, and will design a mapping from one to the other, so as to allow for a translation of one linguistic resource into the other. A typical example of such an approach is POS tagset mapping, or work on the transfer of annotations (for example, across data in different languages). If the linguistic classifications underlying both resouces are isomorphic, transfer is simply a matter of reformatting. Otherwise, mapping rules or conversion routines with potentially complex conditions must be applied.

Alternatively, the *interlingua philosophy on interoperability* analyses both representations at hand, and then constructs a third representa-

tion that is a generalization over both; or it may relate both representations via an ontology of the targeted linguistic description. Thus, the interlingua provides an abstraction over the individual representations to be merged or compared. It may be partly underspecified, leaving room for more descriptive granularity in the individual representations.

Papillon is a typical example of the interlingua philosophy of interoperability. LMF provides another example; it takes the route of underspecification up to the level where the actual LMF standard is rather a meta-specification (i.e. one according to which lexical specifications can be built). Moreover, LMF foresees different solutions as alternatives, whenever relevant (see above, the case of multilingual dictionaries) and thus leaves open options for different approaches and/or different granularity. One of the earliest proposals in line with the interlingua philosophy was the EAGLES proposal for pos-tagset standardization (Leech and Wilson, 1996). Created by means of explicit search for the common denominator of different approaches and different existing tagsets, the EAGLES standard consists of a (small) obligatory core specification and numerous optional refinements. In the core specification, only those descriptive dimensions and values that can safely be assumed to be valid for the languages covered by the EAGLES proposal are included. All others are part of the extensions.

Approaches in line with the interlingual philosophy tend to rely on agreed inventories of descriptive devices, i.e. of meta-specifications (see Figure 1) for data categories. An example for such a meta specification, relating treebank models, is (Sasaki et al., 2003). An example for data category definitions is ITS ("Internationalization Tag Set") (Lieske and Sasaki, 2007), which uses the data categories for inter-relating inventories for localization and internationalization purposes. Linguistic upper ontologies, like GOLD (Farrar and Langendoen, 2003) etc. provides more general and broad inventories for linguistic description. In the course of the last few years, the ISO 12620 standards proposal has been elaborated, which makes recommendations for the creation and maintenance of a Data Category Repository, i.e. an inventory of data categories for linguistic description (Marc Kemps-Snijders and Wright, 2008).

In all cases, the objective is to avoid redundant "synonymous" data categories, and to ease the interlingual common representation of data descriptions from different sources by using data categories from a common inventory. In the ISO 12620 proposal, the categories themselves are not defined, but rather the procedures to be followed in order to include data category proposals in the registry are provided. A group of experts who know the current registry contents in detail decide jointly

whether a given proposal is new or not and therefore whether it should be added to the Registry.

4.3. Linguistic formalisms and interoperability

In the previous sections, several formalisms have been mentioned, especially in the context of resouce architectures. As noted above, the use of formalisms for linguistic representation plays an important role for the technical feasibility of interoperability. This section focusses this very aspect of formalisms. Most formalisms used in computational linguistics are in some sense graph-based, as they mostly rely on attribute-value pairs. All attribute-value pairs can also be expressed by attribute-value graphs (Carpenter, 1992). There are differences, however, as to the generality and, conversely, the role of a formal semantics of these formalisms.

Among the most general graph-based formalisms is the DATR (Evans and Gazdar, 1996) formalism, which consists only of attribute-value pairs. Unification-based processing typically relies on directed acyclic graphs, as for example does Lexical Functional Grammar (Kaplan and Bresnan, 1982).

Work on the Semantic Web has also led to the creation of means for consistency control. The Resource Description Framework (RDF/RDFS) introduces the notion of classes and subclasses, i.e. the possibility to define type hierarchies. For resource modelling, this provides the formal devices to hierarchically organize linguistic knowledge, to introduce different levels of granularity in abstractions used to model linguistic descriptions, and to use simple (is-a) inferences when searching data, (cf. Görz, prep); this inventory of formal properties is relatively similar to that obtained in the 1990s by means of typed feature logic (Carpenter, 1992), for example in HPSG (Pollard and Sag, 1994). RDFS also introduces restrictions, properties and ranges of attribute values, thereby providing roughly the same functionality as HPSG's appropriateness constraint: one can state which descriptive categories are appropriate for a linguistic object of a certain kind, and which values a given attribute may take. Finally, the Web Ontology Language, OWL, (McGuinness and v. Harmelen, 2003), and especially its description logic version (Baader et al., 2003), OWL-DL, makes inferencing possible, as it allows the user to formulate logical constraints over classses and properties. It thus provides the highest level of consistency control, and the freedom, for example, to formally characterize and distinguish different kinds of relations between linguistic objects.

With respect to interoperability, there is thus a trade-off between generality and formal control. More general frameworks, such as those

based on general graph models, offer little control over the data model; however, they support cohabitation of - possibly heterogeneous - data from different sources. Conversely, highly constrained frameworks, such as e.g. OWL-DL, make it easier to create structured formalized representations and to use their formal properties in queries, e.g. through inferencing; however, merging data from heterogeneous sources requires extra effort.

In this trade-off, currently most of the interlingua-oriented models opt for generality over control. Examples are Lexical Systems (Polguère, this issue), Trippel's proposal for a generic lexicon formalism (Trippel, 2006), or the LMF proposal (Francopoulo et al., 2006a). The same holds for standardization proposals concerning the principles of linguistic modelling of static text-based resources, such as the ongoing work on LAF, the *Linguistic Annotation Framework*, on GrAF (Ide and Suderman, 2007), the *Graph Annotation Framework*, and *Annotation Graphs* (Bird and Liberman, 2001). For the field of corpus annotation, the medium term view is that GrAF will provide a general graph-based metamodel for the technical realization of annotations. LAF specifies this in terms of modelling methods, leaving open the details of any specific model of annotated corpora. Such models should likely be specific to the levels of linguistic description in question; in fact, proposals for the level of morphosyntax (tokenizing, POS-tagging) are being made in the MAF metamodel (*Morphosyntactic Annotation Framework*). Moreover, the Syntactic Annotation Framework (SynAF) and the Semantic Annotation Framework (SemAF) address their respective levels of description. These specifications are beeing made within the ISO committee Language Resource Management (TC-37/SC-4). All of these approaches provide a syntax for the representation of potentially annotated linguistic data. A formal semantic interpretation, however, has to be taken from additional devices.

Obviously, such representations allow for a cohabitation of descriptions from different sources, which may follow different approaches, theories, etc. They allow for a very general level of interoperability. On the other hand, they may require specific external interpretation techniques, to allow for an adequate reuse of data represented in a general common format.

5. Conclusion

Over the past few years, resource interoperability has become a major research area, addressing a pressing need of the NLP community. It integrates earlier and parallel work on resource building, standardization

proposals, and formalisms and tools. It is therefore timely to gather together in this special issue a series of focused contributions dedicated to resouce interoperability. The editors would like to thank all authors for their contributions to this publication; they wish to thank in particular the editors of the *Language Resources and Evaluation* journal for providing a well-known platform to make interoperability research known to a wide community of interested experts.

References

Baader, F., D. Calvanese, D. L. /McGuinness, D. Nardi, and P. F. Patel-Schneider (eds.): 2003, *The Description Logic Handbook: Theory, Implementation and Applications*. CUP.

Bird, S. and M. Liberman: 2001, 'A Formal Framework for Linguistic Annotation'. *Speech Communication* **33**(1,2), 23–60.

Boitet, C., M. Mangeot, and G. Sérasset: 2002, 'The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons'. In: *NLPXML '02: Proceedings of the 2nd workshop on NLP and XML*. Morristown, NJ, USA, Association for Computational Linguistics.

Burnard, L. and S. Bauman (eds.): 2007, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

Calzolari, N., A. Zampolli, and A. Lenci: 2002, 'Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative'. In: *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*. London, UK, pp. 264–279, Springer-Verlag.

Carletta, J., S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann: 2003, 'The NITE XML Toolkit: Flexible annotation for multi-modal language data'. *Behavior Research Methods, Instruments, and Computers* **35**(3), 353–363.

Carpenter, B.: 1992, *The Logic of Typed Feature Structures: With Applications to Unification Grammars, Logic Programs and Constraint Resolution*, No. 24 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Cunningham, H.: 2002, 'GATE, a General Architecture for Text Engineering'. *Computers and the Humanities* **36**, 223–254.

Emerson, T. and J. O'Neil: 2006, 'Large Corpus Construction for Chinese Lexicon Development'. In: *Proceedings of the 29th Unicode Conference*. San Francisco, USA.

Evans, R. and G. Gazdar: 1996, 'DATR: A language for lexical knowledge representation'. *Computational Linguistics* **22**(22), 167–216.

Farrar, S. and D. T. Langendoen: 2003, 'A linguistic ontology for the Semantic Web'. *GLOT International* **7**(3), 97–100.

Ferrucci, D. and A. Lally: 2004, 'UIMA: an architectural approach to unstructured information processing in the corporate research environment'. *Nat. Lang. Eng.* **10**(3-4), 327–348.

Francopoulo, G., N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria: 2006a, 'Lexical Markup Framework (LMF) for NLP Multilingual Resources'. In: *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*. Sydney, Australia, pp. 1–8, Association for Computational Linguistics.

Francopoulo, G., M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria: 2006b, 'LMF for multilingual, specialized lexicons'. In: E. Hinrichs, N. Ide, M. Palmer, and J. Pustejovsky (eds.): *Proceedings of the LREC 2006 Satellite Workshop on Merging and Layering Linguistic Information*. Genoa, Italy.

Görz, G.: in prep., 'Representing Computational Dictionaries in AI-Oriented Knowledge Representation Formalisms'. In: *Dictionaries. An International Handbook of Lexicography – Supplementary volume: New developments in lexicography, with a special focus on computational lexicography*, HSK – Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: W. de Gruyter, pp. 10–19. to appear in 2009.

Helbig, H.: 2001, *Die semantische Struktur natürlicher Sprache: Wissensrepräsentation mit MultiNet*. Berlin: Springer.

Ide, N. and K. Suderman: 2007, 'GrAF: A Graph-based Format for Linguistic Annotations'. In: *Proceedings of the ACL Workshop on Linguistic Annotation*. Prague, Czech Republic, pp. 1–8.

ISO 24610-1:2006: 2006, 'Language Resource Management – Feature Structures – Part 1: Feature Structure Representation'. Technical report, International Organization for Standardization.

Kaplan, R. M. and J. Bresnan: 1982, 'Lexical-Functional Grammar: A Formal System for Grammatical Representation'. In: J. Bresnan (ed.): *The Mental Representation of Grammatical Relations*. Cambridge, Massachusetts: MIT Press, pp. 173–281.

Leech, G. and A. Wilson: 1996, 'EAGLES. Recommendations for the Morphosyntactic Annotation of Corpora'. Technical report, Expert Advisory Group on Language Engineering Standards. EAGLES Document EAG-TCWG-MAC/R.

Lieske, C. and F. Sasaki: 2007, 'Internationalization Tag Set (ITS) 1.0. W3C Recommendation'. Technical report, World Wide Web Consortium.

Lyding, V., E. Chiocchetti, G. Sérasset, and F. Brunet-Manquat: 2006, 'The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated'. In: A. Witt, G. Sérasset, S. Armstrong, J. Breen, U. Heid, and F. Sasaki (eds.): *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*. Sydney, Australia, pp. 25–31, Association for Computational Linguistics.

Marc Kemps-Snijders, Menzo Windhouwer, P. W. and S. E. Wright: 2008, 'ISOcat: Corralling Data Categories in the Wild'. In: European Language Resources Association (ELRA) (ed.): *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

McGuinness, D. L. and F. v. Harmelen: 2003, 'OWL Web Ontology Language Overview'. Technical report, W3C. http://www.w3.org/TR/owl-features/.

Pollard, C. and I. A. Sag: 1994, *Head-Driven Phrase Structure Grammar*. Chicago, Illinois: The University of Chicago Press.

Resnik, P., M. B. Olsen, and M. Diab: 1999, 'The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues''. *Computers and the Humanities* **33**(1-2), 129–153.

Richardson, L. and S. Ruby: 2007, *RESTful Web Services*. O'Reilly.

Sasaki, F., A. Witt, and D. Metzing: 2003, 'Declarations of Relations, Differences and Transformations between Theory-specific Treebanks: A New Methodology'. In: J. Nivre (ed.): *The Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*. Vaxjö University, Sweden.

Schäfer, U.: 2006, 'Middleware for Creating and Combining Multi-dimensional NLP markup'. In: *Proceedings of the EACL-2006 Workshop on Multi-dimensional Markup in Natural Language Processing*. Trento, Italy.

Tognini-Bonelli, E.: 2001, *Corpus Linguistics at Work*, Vol. 6 of *Studies in Corpus Linguistics*. Amsterdam: Benjamins.

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

Váradi, T., S. Krauwer, P. Wittenburg, M. Wynne, and K. Koskenniemi: 2008, 'CLARIN: Common Language Resources and Technology Infrastructure'. In: European Language Resources Association (ELRA) (ed.): *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

Witt, A., G. Rehm, E. Hinrichs, T. Lehmberg, and J. Stegmann: 2009, 'SusTEInability of Linguistic Resources through Feature Structures'. *Literary and Linguistic Computing*. In print.

Witt, A., G. Sérasset, S. Armstrong, J. Breen, U. Heid, and F. Sasaki (eds.): 2006, *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*. Sydney, Australia: Association for Computational Linguistics.

Wörner, K., A. Witt, G. Rehm, and S. Dipper: 2006, 'Modelling Linguistic Data Structures'. In: B. T. Usdin (ed.): *Proceedings of Extreme Markup Languages 2006*. Montréal, Canada.