Computational Language Systems: Architectures

H Cunningham and K Bontcheva, University of Sheffield, Sheffield, UK

Software Architecture

Every building, and every computer program, has an 'architecture': structural and organizational principles that underpin its design and construction. The garden shed once built by one of the authors had an *ad hoc* architecture, extracted (somewhat painfully) from the imagination during a slow and nondeterministic process that, luckily, resulted in a structure that keeps the rain on the outside and the mower on the inside (at least for the time being). As well as being *ad hoc* (i.e., not informed by analysis of similar practice or relevant science or engineering) this architecture is implicit: no explicit design was made, and no records or documentation were kept of the construction process. The pyramid in the courtyard of the Louvre, by contrast, was constructed in a process involving explicit design performed by qualified engineers with a wealth of theoretical and practical knowledge of the properties of materials, the relative merits and strengths of different construction techniques, and the like.

So it is with software: sometimes it is thrown together by enthusiastic amateurs, and sometimes it is architected, built to last, and intended to be 'not something you finish, but something you start' (to paraphrase Brand, 1994).

Several researchers argued in the early and middle 1990s that the field of computational infrastructure or architecture for human language computation merited increased attention. The reasoning was that the increasingly large-scale and technologically significant nature of language processing science was placing increasing burdens of an engineering nature on research and development (R&D) workers seeking robust and practical methods (as was the increasingly collaborative nature of research in this field, which puts a large premium on software integration and interoperation). Since then, several significant systems and practices have been developed in what may be called software architecture for language engineering (SALE).

Language engineering (LE) may be defined as the production of software systems that involve processing human language with quantifiable accuracy and predictable development resources (Cunningham, 1999). LE is related to but distinct from the fields of computational linguistics, natural language processing, and artificial intelligence, with its own priorities and concerns. Chief among these are (1) dealing with large-scale tasks of practical utility, (2) measuring progress quantitatively relative to performance on examples of such tasks, (3) a growing realization of the importance of software engineering in general, and (4) reusability, robustness, efficiency, and productivity, in particular. Software architectures can contribute significantly toward achieving these goals (Maynard *et al.*, 2002; Cunningham and Scott, 2004).

This article gives a critical review of the various approaches that have been taken to the problem of software architecture for language engineering (SALE). The prime criterion for inclusion in this article is that the approaches are *infrastructural* – work that is intended to support language engineering (LE) R&D in some way that extends beyond the boundaries of a single time-limited project.

This article presents categories of work that range over a wide area. To provide an organizing principle for the discussion, we extrapolate a set of architectural issues that represent the union of those addressed by the various researchers cited. This approach has the advantage of making it easier to see how certain problems have been addressed and the disadvantage that multipurpose infrastructures appear in several categories.

The following section discusses infrastructures aimed at algorithmic resources including the issues of component integration and execution. The article then analyzes data resources infrastructure, including the issues of access and the representation of information about text and speech. If concludes with a discussion on future directions for work on SALE.

Software Architectures for Language Engineering

The problem addressed by the systems reviewed here is the construction of software infrastructure for language processing: software that is intended to apply to whole families of problems within this field and to be like a craftsman's toolbox in the service of construction and experimentation. We consider three types of infrastructural systems: frameworks, architectures, and development environments.

A 'framework' typically means an object-oriented class library that has been designed with a certain domain in mind and that can be tailored and extended to solve problems in that domain. A framework may also be known as a platform or a component system.

All software systems have an architecture. Sometimes, the architecture is explicit, perhaps conforming to certain standards or patterns, and sometimes it is implicit. Where an architecture is explicit and targeted on more than one system, it is known as a 'reference architecture' or a 'domain-specific architecture.' The former is "a software architecture for a family of application systems" (Tracz and Mar, 1995). The term 'domain-specific software architecture (DSSA),' the subject of an eponymous ARPA research program, "applies to architectures designed to address the known architectural abstractions specific to given problem domains" (Clements and Northron, 1996).

An implementation of an architecture that includes some graphical tools for building and testing systems is a 'development environment'. One of the benefits of an explicit and repeatable architecture is that it can give rise to a symbiotic relationship with a dedicated development environment. In this relationship, the development environment can help designers conform to architectural principles and visualize the effect of various design choices and can provide code libraries tailored to the architecture.

The most significant issues addressed by SALE systems include the following.

- enabling a clean separation of low-level tasks, such as data storage, data visualization, location and loading of components, and execution of processes from the data structures and algorithms that actually process human language
- reducing integration overheads by providing standard mechanisms for components to communicate data about language and using open standards, such as Java and XML, as the underlying platform
- providing a baseline set of language processing components that can be extended and/or replaced by users as required
- providing a development environment or at least a set of tools to support users in modifying and implementing language processing components and applications
- automating measurement of performance of language-processing components.

This article focuses on the first two sets of issues, because they are issues that arise in every single NLP system or application and are prime areas where SALE can make a contribution. For a discussion of other requirements, see Cunningham (2000).

Categories of Work on SALE

As with other software, LE programs comprise data and algorithms. The current trend in software development is to model both data and algorithms together, as 'objects.' (Older development methods, such as structured analysis kept them largely separate; Yourdon, 1989.) Systems that adopt the new approach are referred to as 'object-oriented' (OO), and there are good reasons to believe that OO software is easier to build and maintain (see Booch, 1994).

In the domain of human language processing R&D, however, the choice is not quite so clear cut. Language data, in various forms, are of such significance in the field that they are frequently worked on independently of the algorithms that process them. Such data have even come to have their own term: 'language resources' (LRs; LREC-1, 1998), covering many data sources, from lexicons to corpora.

In recognition of this distinction, this article uses the following terminology.

- **Language resource** (**LR**) refers to data-only resources, such as lexicons, corpora, thesauri, or ontologies. Some LRs come with software (e.g., Wordnet has both a user query interface and C and Prolog APIs), but resources in which software is only a means of accessing the underlying data are still defined as LRs.
- **Processing resource** (**PR**) refers to resources that are principally programmatic or algorithmic, such as lemmatizers, generators, translators, parsers, or speech recognizers. For example a part-of-speech (POS) tagger is best characterized by reference to the process it performs on text. PRs typically include LRs (e.g., a tagger often has a lexicon).

PRs can be viewed as algorithms that map between different types of LR and that typically use LRs in the mapping process. An MT (Machine Translation) engine, for example, maps a monolingual corpus into a multilingual aligned corpus using lexicons, grammars, and the like.

Adopting the PR/LR distinction is a matter of conforming to established domain practice and terminology. It does not imply that one cannot model the domain (or build software to support it) in an object-oriented manner. This distinction is used to categorize work on SALE. The next section surveys infrastructural work on processing resources, and the following section reviews the much more substantial body of work on language resources.

Processing Resources

Often, a language processing system follows several discrete steps. For example, a translation application must first analyze the source text to arrive at some representation of meaning before it can begin deciding upon target language structures that parallel that meaning. A typical language analysis process follows such stages as text structure analysis, tokenization, morphological analysis, syntactic parsing, and semantic analysis. The exact breakdown varies widely and is to some extent dependent on method; some statistical work early in the second wave of the application of these types of method completely ignored the conventional language analysis steps in favor of a technique based on a memory of parallel texts (Brown *et al*., 1990). Later work has tended to accept the advantages of some of these stages, however, though they may be moved into an off-line corpus annotation process, such as the Penn Treebank (Marcus *et al*., 1993).

Each of these stages is represented by components that perform processes on text and use components containing data about language, such as lexicons and grammars. In other words, the analysis steps are realized as a set of processing resources (PRs). Several architectural questions arise in this context:

1. Is the execution of the PRs best done serially or in parallel?
2. How should PRs be represented such that their discovery on a network and loading into an executive process are transparent to the developer of their linguistic functions?
3. How should distribution across different machines be handled?
4. What information should be stored about components, and how should it be represented?
5. How can commonalities among component sets be exploited?
6. How should the components communicate information between each other? (This question can also be stated as, 'How should information about text and speech be represented?')

This section reviews work that addresses questions 1–5. The issue of representing information about language is addressed in the following section.

Locating and Loading

There are several reasons why PR components should be separate from the controlling application that executes them:

- There will often be a many-to-one relation between applications and PRs. Any application using language analysis technology needs a tokenizer component, for example.
- A PR may have been developed for one computing platform, such as UNIX, but the application wishing to use it may operate on another (e.g., Windows).
- The processing regime of the application may require linear or asynchronous execution; this choice should be isolated from the component structures as far as possible to promote generality and encourage reuse.
- PR developers should not be forced to deal with application-level software engineering issues, such as how to manage installation, distribution over networks, exception handling, and so on.
- Explicit modeling of components allows exploitation of modern component infrastructures, such as Java Beans or Active X.

Accordingly, many papers on infrastructural software for LE separate components from the control executive (e.g., Boitet and Seligman, 1994; Edmondson and Iles, 1994; Koning *et al*., 1995; Wolinski *et al*., 1998; Poirier, 1999; Zajac, 1998b; Lavelli *et al*., 2002; Cunningham *et al*., 2002a). The term 'executive' is used here in the sense of a software entity that executes, or runs, other entities. The questions then are how do components become known to control processes or applications and how are they loaded and initialized. A related question is what data should be stored with components to facilitate their use by an executive; see the discussion on metadata below. Much work ignores component-related issues the rest of this section covers those SALE systems for which the data are available.

The TIPSTER architecture (Grishman, 1997) recognized the existence of the locating and loading problems, but did not provide a full solution to the problem. The architecture document includes a placeholder for such a solution – in the form of a 'register annotator' Application Programmers' Interface (API) call, which an implementation could use to provide component loading – but the semantics of the call were never specified.

The TalLab architecture "is embedded in the operating system," which allows them to "reuse directly a huge, efficient and reliable amount of code" (Wolinski *et al*., 1998). The precise practicalities of this choice are unclear, but it seems that components are stored in particular types of directory structure, which are presumably known to the application at startup time.

The Intarc Communication Environment (ICE) is an "environment for the development of distributed AI systems" (Amtrup, 1995) and part of the Verbmobil real-time speech-to-speech translation project (Kay *et al*., 1994). ICE provides distribution based around Parallel Virtual Machine (PVM) and a communication layer based on channels. ICE is not specific to LE because the communication channels do not use data structures specific to NLP needs and because document handling issues are left to the individual modules. ICE's answer to the locating and loading problem is the Intarc License Server, which is a kind of naming service or registry that stores addressing information for components. Components must themselves register with the server by making an API call (Ice_Attach). The components must therefore link to the ICE libraries and know the location of the license server as must applications using ICE services.

Following from the ICE work, Herzog *et al.* (2004) presented the latest in three generations of architecture to arise from the Verbmobil and Smartkom projects, in the shape of the Multiplatform system. This architecture supports multiple distributed components from diverse platforms and implementation languages running asynchonously and communicating via a message-passing substrate.

Corelli (Zajac, 1997) and its successor, Calypso, (Zajac, 1998b) are also distributed systems that cater for asynchronous execution. The initial Corelli system implemented much of the CORBA standard (Object Management Group, 1992), and component discovery used a naming and directory service. All communication and distribution were mediated by an object request broker (ORB). Components ran as servers and implemented a small API to allow their use by an executive or application process. In the later Calypso incarnation, CORBA was replaced by simpler mechanisms because efficiency problems (for a usage example, see Amtrup, 1999). In Calypso, components are stored in a centralized repository, which sidesteps the discovery problem. Loading is handled by requiring components to implement a common interface.

Another distributed architecture based on CORBA is SiSSA (Lavelli *et al.*, 2002). The architecture comprises processors (PRs in our terms), servers for their execution, data containers (LRs), and a manager component called SiSSA Manager, which establishes and removes connections between the processors, according to a user-designed data flow. SiSSA uses a processor repository to keep information about processors registered with the architecture.

Carreras and Padró (2002) reported a distributed architecture specifically for language analyzers.

GATE version 1 (Cunningham *et al.*, 1997) was a single-process, serial execution system. Components had to reside in the same file system as the executive; location was performed by searching a path stored in an environment variable. Loading was performed in three ways, depending on the type of component and which of the GATE APIs it used.

GATE version 2 (Cunningham *et al.*, 2002a,b) supports remote components; location is performed by providing one or more component repositories called Collection of REusable Objects for Language Engineering (CREOLE) repositories, which contain XML definitions of each resource and the types of its parameters (e.g., whether it works with documents or corpora). The user can then instantiate a component by selecting it from the list of available components and choosing its load-time parameters. GATE makes a distinction between load-time and run-time parameters; the former are essential for the working of the module (e.g., a grammar) and need to be provided at load time, whereas the latter can change from one execution to the next (e.g., a document to be analyzed). Components can also be re-initialized, which enables users to edit their load-time data (e.g., grammars) within the graphical environment and then reload the component to reflect the changes. GATE also supports editing of remote language resources and execution of remote components using remote method invocation (RMI); that is, it provides facilities for building client-server applications.

Execution

It seems unlikely that people process language by means of a set of linear steps involving morphology, syntax, and so on. More likely, we deploy our cognitive faculties in a parallel fashion; hence, the term 'parallel distributed processing' in neural modeling work (McClelland and Rumelhart, 1986). These kinds of ideas have motivated work on nonlinear component execution in NLP; von Hahn (1994) gave an overview of a number of approaches, and a significant early contribution was the Hearsay speech understanding system (Erman *et al.*, 1980).

Examples of asynchronous infrastructural systems include Kasuga (Boitet and Seligman, 1994), Pantome (Edmondson and Iles, 1994), Talisman (Koning *et al.*, 1995), Verbmobil (Görz *et al.*, 1996), TalLab (Wolinski *et al.*, 1998), Xelda (Poirier, 1999), Corelli (Zajac, 1997), Calypso (Zajac, 1998b), SiSSA (Lavelli *et al.*, 2002), Distributed Inquery (Cahoon and McKinley, 1996), and the Galaxy Communicator Software Infrastructure (GCSI-MITRE, 2002). Motivations include the desire for nonlinear execution and for feedback loops in ambiguity resolution (see Koning *et al.*, 1995).

In the Inquery and Verbmobil systems, an additional motivation is efficiency. ICE, the Verbmobil infrastructure, addressed two problems: distributed processing and incremental interpretation. Distribution is intended to contribute to processing speed in what is a very computer-intensive application area (speech-to-speech translation). Incremental interpretation is designed both for speed and to facilitate feedback of results from downstream modules to upstream ones (e.g., to inform the selection of word interpretations from phone lattices using POS information). ICE's PVM-based architecture provides for distributed asynchronous execution.

GCSI is an open source architecture for constructing dialogue systems. This infrastructure concentrates on distributed processing, hooking together sets of servers and clients that collaborate to hold dialogues with human interlocutors. Data get passed between these components as attribute/value sets or 'frames,' the structuring and semantics of which must be agreed upon on a case-by-case basis. Communication between modules is achieved using a hub. This architectural style tends to treat components as black boxes that are developed using other tool sets. To solve this problem, other support environments can be used to produce GCSI server components, using GCSI as a communication substrate to integrate with other components.

The model currently adopted in GATE is that each PR may run in its own thread if asynchronous processing is required (by default, PRs will be executed serially in a single thread). The set of LRs being manipulated by a group of multithreaded PRs must be synchronized (i.e., all their methods must have locks associated with whichever thread is calling them at a particular point). Synchronization of LRs is performed in a manner similar to the Java collections framework. This arrangement allows the PRs to share data safely. Responsibility for the semantics of the interleaving of data access (who has to write what in what sequence for the system to succeed) is a matter for the user, however.

Metadata

A distinction may be made between the data that language processing components use (or language resources) and data that are associated with components for descriptive and other reasons. The latter are sometimes referred to as 'metadata' to differentiate them from the former. In a similar fashion web content is largely expressed in HTML; data that *describe* web resources, such as 'this HTML page is a library catalogue,' are also called metadata. Relevant standards in this area include the Resource Description Framework RDF; (Lassila and Swick, 1999; Berners-Lee *et al.*, 1999).

There are several reasons why metadata should be part of a component infrastructure, including the following:

- to facilitate the interfacing and configuration of components
- to encode version, author, and availability data
- to encode purpose data and allow browsing of large component sets.

When components are reused across more than one application or research project, often their input/output (I/O) characteristics have not been designed alongside the other components forming the language-processing capability of the application. For example, one POS tagger may require tokens as input in a one-per-line encoding. Another may require the Standard Generalized Markup Language (SGML) input (Goldfarb, 1990). To reuse the tagger with a tokenizer that produces some different flavor of output, that output must be transformed to suit the tagger's expectations. In cases where there is an isomorphism between the available output and the required input, a straightforward syntactic mapping of representations is possible. In cases where there is a semantic mismatch, additional processing is necessary.

Busemann (1999) addressed component interfacing and described a method for using feature structure matrices to encode structural transformations on component I/O data structures. These transformations essentially reorder the data structures around pre-existing unit boundaries; therefore, the technique assumes isomorphism among the representations concerned. The technique also allows for type checking of the output data during restructuring.

TIPSTER (Grishman, 1997), GATE (Cunningham, 2002), and Calypso (Zajac, 1998b) deal with interfacing in two ways. First, component interfaces share a common data structure (e.g., corpora of annotated documents), thus ensuring that the syntactic properties of the interface are compatible. Component wrappers are used to interface to other representations as necessary; for example, a Brill tagger (Brill, 1992) wrapper writes out token annotations in the required one-per-line format, then reads in the tags, and writes them back to the document as annotations. Second, where there is semantic incompatibility between the output of one component and the input of another, a dedicated transduction component can be written to act as an intermediary between the two.

In Verbmobil a component interface language is used, which constrains the I/O profiles of the various modules (Bos *et al.*, 1998). This language is a Prolog term that encodes logical semantic information in a flat list structure. The principle is similar to that used in TIPSTER-based systems, but the applicability is somewhat restricted by the specific nature of the data structure.

Provision of descriptive metadata has been addressed by the Natural Language Software Registry (NLSR; DFKI, 1999) and by the EUDICO distributed corpora project (Brugman *et al.*, 1998a,b). In each case, web-compatible data (HTML and XML, respectively) are associated with components. The NLSR is purely a browsable description; the EUDICO work links the metadata with the resources themselves, allowing the launching of appropriate tools to examine them. Note that EUDICO has only dealt with language resource components to date. GATE 2 (Cunningham *et al.*, 2002b) uses XML for describing the metadata associated with processing resources in its CREOLE repositories. This metadata are used for component loading and also for launching the corresponding visualization and editing tools.

In addition to the issue of I/O transformation, in certain cases it may be desirable to be able to identify automatically which components are plug-compatible with which other ones, so as to identify possible execution paths through the component set.

GATE 1 (Cunningham *et al.*, 1997) addresses automatic identification of execution paths by associating a configuration file with each processing component that details the input (preconditions) and output (post-conditions) in terms of TIPSTER annotation and attribute types (see the section on reference attribution). This information is then used to auto-generate an execution graph for the component set.

Commonalities

To conclude this survey of infrastructural work related to processing, this section looks at the exploitation of commonalities between components. For example, both parsers and taggers have the characteristics of language analyzers. One of the key motivating factors for SALE is to break the 'software waste cycle' (Veronis and Ide, 1996) and promote reuse of components. Various researchers have approached this issue by identifying typical component sets for particular tasks (Hobbs, 1993; TIPSTER, 1995; Reiter and Dale, 2000). Work is continuing on providing implementations of common components (Ibrahim and Cummins, 1989; Cheong *et al.*, 1994). The rest of this section describes these approaches.

Reiter and Dale have reviewed and categorized Natural Language Generation (NLG) components and systems in some detail. Reiter (1994) argued that a consensus component breakdown has emerged in NLG (and that there is some psychological plausibility for this architecture); the classification was extended in Reiter and Dale (2000). They also discussed common data structures in NLG (as does the RAGS project; see below) and appropriate methods for the design and development of NLG systems. Reiter (1999) argued that the usefulness of this kind of architectural description is to 'make it easier to describe functionalities and data structures' and thus facilitate research by creating a common vocabulary among researchers. He stated that this is a more limited but more realistic goal than supporting the integration of diverse NLG components in an actual software system. The term he used for this kind of descriptive work is a 'reference architecture,' which is also the subject of the workshop at which the paper was presented (Mellish and Scott, 1999).

The TIPSTER research program developed descriptive or reference architectures for information extraction and for information retrieval. Hobbs (1993) described a typical module set for an IE system. The architecture comprises 10 components, dealing with such tasks as pre-processing, parsing, semantic interpretation, and lexical disambiguation; for a description of the full set, see Gaizauskas and Wilks, 1998). For IR, TIPSTER (1995) describes two functions, search and routing, each with a typical component set (some of which are PRs and some LRs.)

An architecture for spoken dialogue systems, which divides the task into dialogue management, context tracking, and pragmatic adaptation, is presented in LuperFoy *et al.* (1998). This in turn leads to an architecture in which various components (realized as agents) collaborate in the dialogue. Some example components are speech recognition, language interpretation, language generation, and speech synthesis. In addition a dialogue manager component provides high-level control and routing of information among components.

The preceding discussion illustrates that there is considerable overlap among component sets developed for various purposes. A SALE that facilitated multipurpose components would cut down on the waste involved in the continual reimplementation of similar components in different contexts. The component model given in Cunningham (2000) is made available in the GATE framework (Cunningham *et al.*, 2002b). This model is based on inheritance: A parser is a type of language analyzer that is a type of processing resource. Language engineers can choose, therefore, between implementing a more specific interface and adhering to the choices made by the GATE developers for that type, or implementing a more general interface and making their own choices about the specifics of their particular resource.

In several cases, work on identifying component commonalities has led to the development of toolkits that aim to implement common tasks in a reusable manner. For example, TARO (Ibrahim and Cummins, 1989) is an OO syntactic analyzer toolkit based on a specification language. A toolkit for building IE systems and exemplified in the MFE IE system is presented in Cheong *et al.* (1994).

Language Resources

As described above, language resources are data components, such as lexicons, corpora, and language models. They are the raw materials of language engineering. This section covers five issues relating to infrastructure for LRs:

1. computational access (local and nonlocal)
2. managing document formats and document collections (corpora), including multilingual resources
3. representing information about corpora (language data or performance modeling)
4. representing information about language (data about language or competence modeling)
5. indexing and retrieval of language-related information.

Note also that the advantages of a component-based model presented (in relation to PRs) in the section on locating and loading PRs also apply to LRs.

Programmatic Access

LRs are of worth only inasmuch as they contribute to the development and operation of PRs and the language processing research prototypes, experiments, and applications that are built from them. A key issue in the use of LRs for language processing purposes is that of computational access. Suppose that a developer is writing a program to generate descriptions of museum catalogue items this program may have a requirement for synonyms, for example, in order to lessen repetition. Several sources for synonyms are available, such as WordNet (Miller, 1990) or Roget's *Thesaurus*. To reuse these sources, the developer needs to access the data in these LRs from their program.

Although the reuse of LRs has exceeded that of PRs (Cunningham *et al.*, 1994), in general, there are still two barriers to LR access and hence LR reuse: (1) each resource has its own representation syntax and corresponding programmatic access mode (e.g., SQL for Celex, C or Prolog for WordNet); and (2) resources must generally be installed locally to be usable, and how this is done depends on what operating systems are available, what support software is required, and the like, which vary from site to site.

A consequence of the first barrier is that, although resources of the same type usually have some structure in common (for example, at one of the most general levels of description, lexicons are organized around words), this commonality cannot be exploited when it comes to using a new resource. In each case, the user has to adapt to a new data structure; this adaptation is a significant overhead. Work that seeks to investigate or exploit commonalities among resources has first to build a layer of access routines on top of each resource. So, for example, if one wished to do task-based evaluation of lexicons by measuring the relative performance of an IE system with different instantiations of lexical resource, one would typically have to write code to translate several different resources into SQL or some other common format. Similarly, work, such as Jing and McKeown (1998) on merging large-scale lexical resources (including WordNet and Comlex) for NLG, must deal with this problem.

There have been two principal responses to this problem: standardization and abstraction. The standardization solution seeks to impose uniformity by specifying formats and structures for LRs. So, for example, the EAGLES working groups have defined standards for lexicons, corpora, and so on (EAGLES, 1999). More recently, Ide and Romary (2004) reported the creation of a framework for linguistic annotations as part of the work of ISO standardization Technical Committee 37, Sub-Committee 4, whose objective

> is to prepare various standards by specifying principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes. These standards will also cover the information produced by natural language processing components in these various domains.

The work reported here is from Working Group 1 of the committee, which has developed a linguistic annotation framework based on the XML (eXtensible Markup Language), RDF(S) (Resource Discovery Framework (Schema)), and OWL (Ontology Web Language).

Although standardization would undoubtedly solve the representation problem, there remains the question of existing LRs (and of competing standards). Peters *et al.* (1998) and Cunningham *et al.* (1998) described experiments with an abstraction approach based on a common object-oriented model for LRs that encapsulates the union of the linguistic information contained in a range of resources and encompasses as many object hierarchies as there are resources. At the top of the resource hierarchies are very general abstractions; at the leaves are data items specific to individual resources. Programmatic access is available at all levels, allowing the developer to select an appropriate level of commonality for each application. Generalizations are made over different object types in the resources, and the object hierarchies are linked at whatever levels of description are appropriate. No single view of the data is imposed on the user, who may choose to stay with the 'original' representation of a particular resource or to access a model of the commonalities among several resources, or a combination of both.

A consequence of the requirement for local installation – the second barrier to LR access – is that users may have to adjust their compute environments to suit resources tailored to particular platforms. In addition, there is no way to 'try before you buy,' no way to examine an LR for its suitability for one's needs before licensing it *in toto*. Correspondingly, there is no way for a resource provider to give limited access to their products for advertising purposes or to gain revenue through piecemeal supply of sections of a resource.

This problem of non local access has also attracted two types of responses, which can be broadly categorized as: web browsing and distributed databases.

Several sites now provide querying facilities from HTML pages, including the Linguistic Data Consortium and the British National Corpus server. So, for example, all occurrences of a particular word in a particular corpus may be found via a web browser. This is a convenient way to access LRs for manual investigative purposes, but is not suited to (or intended for) use by programs for their access purposes.

Moving beyond browsing, several papers report work on programmatic access using distributed databases. Fikes and Farquhar (1999) showed how ontologies may be distributed, Brugman *et al.* (1998a,b) described the EUDICO distributed corpus access system, and Peters *et al.* (1998) and Cunningham *et al.* (1998) proposed a system similar to EUDICO, generalized to other types of LR. Some new directions in sharing language resources are discussed in the section on trends.

Other issues in the area of access to LRs include that of efficient indexing and search of corpora (see the section, 'Indexing and Retrieval'), and that of annotation of corpora (see the section on annotation). The issue of how to access SGML documents in an efficient manner is discussed in Olson and Lee (1997), who investigated the use of object-oriented databases for storing and retrieving SGML documents. Their conclusions were essentially negative due to the slowness of the databases used. Hendler and Stoffel (1999) discussed how ontologies may be stored and processed efficiently using relational databases, and here the results were more positive.

Documents, Formats, and Corpora

Documents play a central role in LE. They are the subject of analysis for such technologies as IE, and they are both analyzed and generated in technologies such as MT. In addition, a large amount of work uses annotated documents as training data for machine learning of numerical models. Previous work on LE infrastructure has developed models for documents and corpora, provided abstraction layers for document formats, and investigated efficient storage of documents in particular formats.

Documents may contain text, audio, video or a mixture of these formats; documents with a mixture of formats are referred to as multimedia documents. The underlying data are frequently accompanied by formatting information (delineating titles, paragraphs, areas of bold text, etc.) and, in the LE context, by annotation (storing linguistic data, such as gesture tags, POS tags, or syntax trees). Both formatting and annotation come in a wide variety of formats, including proprietary binary data, such as MS Word's.doc or Excel's .xls; semi-open, semi-readable formats, such as Rich Text Format (Word's exchange format); and nonproprietary standardized formats, such as HTML, XML, or GIF (Graphics Interchange Format).

The Text Encoding Initiative (TEI; (Sperberg-McQueen and Burnard, 1994, 2002), the Corpus Encoding Standard (CES; Ide, 1998), and XCES (Ide *et al.*, 2000) are models of documents and corpora that aim to standardize the representation of structural and linguistic data for textual documents. The general approach is to represent *all* information about document structure, formatting, and linguistic annotation using SGML/XML.

The issue of document formats has been addressed by several TIPSTER-based systems, including GATE and Calypso, and by the HTK speech recognition toolkit (Young *et al.*, 1999). In the HTK toolkit, the approach is to provide API calls that deal with documents in various known formats (e.g. Windows audioformat, MPEG) independent of those formats. For example, a speech recognizer can access the raw audio from these documents without knowing anything about the representation format.

The TIPSTER systems deal with formats by means of input filters that contain knowledge about the format encoding and use that knowledge to unpack format information into annotations. TIPSTER also supplies a model of corpora and data associated with both corpus and documents (Grishman, 1997). Note that the two approaches are not mutually exclusive: Ogden (1999) has defined a mapping between TEI/CES and TIPSTER annotations.

Another important issue that needs to be dealt with in infrastructures supporting LRs in multiple languages is the problem of editing and displaying multilingual information. It is often thought that the character sets problem has been solved by use of the Unicode standard. This standard is an important advance, but in practice the ability to process text in a large number of the world's languages is still limited by (1) incomplete support for Unicode in operating systems and applications software, (2) languages missing from the standard, and (3) difficulties in converting non-Unicode character encodings to Unicode. To deal with all these issues, including displaying and editing of Unicode documents, GATE provides a Unicode Kit and a specialized editor (Tablan *et al.*, 2002). In addition, all processing resources and visualization components are Unicode-compliant.

Annotation

One of the key issues for much of the work done in this area is how to represent information about text and speech. This kind of information is sometimes called 'language data,' distinguishing it from 'data about language' in the form of lexicons, grammars, etc.

Two broad approaches to annotation have been taken: to use markup (e.g., SGML/XML) or to use annotation data structures with references or pointers to the original (e.g., TIPSTER, ATLAS). Interestingly, the differences between the two kinds of approaches have become less pronounced in recent work. SGML used to involve embedding markup in the text; TIPSTER (and related systems) use a referential scheme where the text remains unchanged and annotation refers to it by character offsets. The embedding approach has several problems, including the difficulty of extending the model to cope with multimedia data (Nelson, 1997, Cunningham *et al.*, 1997; Bird and Liberman, 1999a). Partly in response to these difficulties and as part of the rebirth of SGML as XML (Goldfarb and Prescod, 1998), the 'ML' community has adopted a referential scheme itself, which is now known as 'stand-off markup.' The data models of the various systems are now much closer than they were before XML existed, and the potential for interoperation between referential systems, such as GATE and XML-based architectures, is greater as a result. GATE exploits this potential by providing input from and output to XML in most parts of the data model (Cunningham *et al.*, 2002a,b).

Markup-Based Architectures Language data can be represented by embedding annotation in the document itself, at least in the case of text documents; users of embedding typically transcribe speech documents before markup or use 'stand-off markup.' The principal examples of embedded markup for language data use the Standard Generalized Markup Language (SGML; Goldfarb, 1990). SGML is a 'meta-language,' a language used to create other languages. The syntax of SGML is therefore abstract, with each document filling in this syntax to obtain a concrete syntax and a particular markup language for that document. In practice, certain conventions are so widespread as to be *de facto* characteristics of SGML itself. For example, annotation is generally delimited by <TAG> and </TAG> pairs, often with some attributes associated, such as <TAG ATTRIBUTE = value>. The legitimate tags (or 'elements') and their attributes and values must be defined for each class of document, using a Document-Type Definition (DTD). It does *not* specify what the markup means; the DTD is the grammar that defines how the elements may be legally combined and in what order in a particular class of text; see Goldfarb (1990). A good example of SGML used for corpus annotation is the British National Corpus (BNC; Burnard, 1995).

The HyperText Markup Language (HTML) is an application of SGML and is specified by its own DTD. A difference from ordinary SGML is that the DTD is often cached with software, such as web browsers, rather than being a separate file associated with the documents that instantiate it. In practice, web browsers have been lenient in enforcing conformance to the HTML DTD, which has led to diversity among web pages; this means that HTML DTDs now represent an idealized specification of the language that often differs from its usage in reality.

Partly in response to this problem, the eXtensible Markup Language (XML; Goldfarb and Prescod, 1998) was developed. SGML is a complex language: DTDs are difficult to write, and full SGML is difficult to parse. XML made the DTD optional and disallowed certain features of SGML, such as markup minimization. For example, the American National Corpus (ANC; Macleod *et al.*, 2002) uses XML and XCES (Ide *et al.*, 2000) to encode linguistic annotations.

One of the problems in the SGML/XML world is that of computational access to and manipulation of markup information. Addressing this problem, the Language Technology group at the University of Edinburgh developed an architecture and framework based on SGML called the LT Normalized SGML Library (LT NSL; McKelvie *et al.*, 1998). This in turn led to the development of LT XML (Brew *et al.*, 1999), following the introduction of the XML standard.

Tools in an LT NSL system communicate via interfaces specified as SGML DTDs (essentially tag set descriptions), using character streams on pipes: a pipe-and-filter arrangement modeled after UNIX-style shell programming. To avoid the need to deal with certain difficult types of SGML (e.g., minimized markup), texts are converted to a normal form before processing. A tool selects what information it requires from an input SGML stream and adds information as new SGML markup. LT XML is an extension of LT NSL to XML; it makes the normalization step unnecessary.

Other similar work in this area includes the XDOC workbench (Rösner and Kunze, 2002), stand-off markup for NLP tools (Artola *et al.*, 2002), and the multilevel annotation of speech (Cassidy and Harrington, 2001).

Reference Annotation I: TIPSTER The ARPA-sponsored TIPSTER program in the United States, which was completed in 1998, produced a data-driven architecture for NLP systems (Grishman, 1997) several sites implemented the architecture, such as GATE version 1 (Cunningham *et al.*, 1999) and ELLOGON (Petasis *et al.*, 2002); the initial prototype was written by Ted Dunning at the Computing Research Lab of New Mexico State University. In contrast to the embedding approach, in TIPSTER, the text remains unchanged while information about it is stored in a separate database. The database refers to the text by means of offsets. The data are stored by reference.

Information is stored in the database in the form of annotations, which associate arbitrary information (attributes) with portions of documents (identified by sets of start/end character offsets or spans). Attributes are often the result of linguistic analysis (e.g., POS tags). In this way, information about texts is kept separate from the texts themselves. In place of an SGML DTD (or XML XSchema), an 'annotation type declaration' defines the information present in annotation sets (though few implementations instantiated this part of the architecture). Figure 1 gives an example of TIPSTER annotation; it "shows a single sentence and the result of three annotation procedures: tokenization with part-of-speech assignment, name recognition, and sentence boundary recognition. Each token has a single attribute, its part of speech (POS); ...; each name also has a single attribute, indicating the type of name: person, company, etc." (Grishman, 1997).

Text

Kevin admired his bike.

0...|5...|10..|15..|20

Annotations

ID	Type	Span		Attributes
		Start	End	
1	token	0	5	pos=NP
2	token	6	13	pos=VBD
3	token	14	17	pos=PP
4	token	18	22	pos=NN
5	token	22	23	
6	name	0	5	name_type=person
7	sentence	0	23	

Figure 1 Example of a TIPSTER annotation.

Documents are grouped into collections (or corpora), each with an associated database storing annotations and such document attributes as identifiers, headlines, etc. The definition of documents and annotations in TIPSTER forms part of an object-oriented model that can deal with inter-as well as intratextual information by means of reference objects that can point at annotations, documents, and collections. The model also describes elements of IE and IR systems relating to their use, providing classes representing queries and information needs.

TIPSTER-style models have several advantages and disadvantages. Texts may appear to be one-dimensional, consisting of a sequence of characters, but this view is incompatible with such structures as tables, which are inherently two-dimensional. Their representation and manipulation are easier in a referential model like TIPSTER than in an embedding one like SGML, in which markup is stored in a one-dimensional text string. In TIPSTER, a column of a table can be represented as a single object with multiple references to parts of the text (an annotation with multiple spans, or a document attribute with multiple references to annotations). Marking columns in SGML requires a tag for each row of the column, and manipulation of the structure as a whole necessitates traversal of all the tags and construction of some other, non-SGML data structure.

Distributed control has a relatively straightforward implementation path in a database-centered system like TIPSTER: the database can act as a blackboard, and implementations can take advantage of well-understood access control technology.

In TIPSTER, in contrast to the hyperlinking used in LT XML, there is no need to break up a document into smaller chunks, as the database management system (DBMS) in the document manager can deal efficiently with large data sets and visualization tools can give intelligible views into this data. To cross-refer between annotations is a matter of citing ID numbers, which are themselves indexes into database records and can be used for efficient data access. It is also possible to have implicit links: Simple API calls find all the token annotations subsumed by a sentence annotation, for example, via their respective byte ranges without any need for additional cross-referencing information.

Another advantage of embedded markup in TIPSTER is that an SGML structure like <w id = p4. w1> has to be parsed in order to extract the fact that there is a 'w' tag whose 'id' attribute is 'p4.w1'. A TIPSTER annotation is effectively a database record with separate fields for type (e.g., 'w'), ID, and other attributes, all of which may be indexed and none of which ever requires parsing.

There are three principal disadvantages of the TIPSTER approach.

1. Editing of texts requires offset recalculation.
2. TIPSTER specifies no interchange format, and TIPSTER data are weakly typed. There is no effective DTD mechanism, though this may also to an extent be an advantage, as a complex typing scheme can inhibit unskilled users.
3. The reference classes can introduce brittleness in the face of changing data: Unless an application chases all references and updates them as the objects they point to change, the data can become inconsistent. This problem also applies to hyperlinking in embedded markup.

Reference Annotation II: Linguistic Data Consortium The Linguistic Data Consortium (LDC) has proposed the use of Directed Acyclic Graphs (DAGs) or just Annotation Graphs (AGs) as a unified data structure for text and speech annotation (Bird *et al.*, 2000b). Bird and Liberman (1999b) provided an example of using these graphs to mark up discourse-level objects. This section compares the structure of TIPSTER annotations with the graph format.

As discussed above, TIPSTER annotations are associated with documents and have four elements:

1. a type, which is a string
2. an ID, which is a string unique among annotations on the document
3. a set of spans that point into the text of the document
4. a set of attributes.

TIPSTER attributes, which are associated with annotations and with documents and collections of documents, have a name, which is a string, and a value, which may be one of several data types including a string; a reference to an annotation, document, or collection; or a set of strings or references. Some implementors of the architecture, including GATE and Corelli, have relaxed the type requirements on attribute values, allowing any object as a value. This has the advantage of flexibility and the disadvantage that it makes viewing, editing, and storage of annotations more complex.

TIPSTER explicitly models references between annotations with special reference classes. These classes rely on annotations, documents, and collections of documents having unique identifiers.

LDC annotations are arcs in a graph, the nodes of which are time points or, by extension, character offsets in a text. Each annotation has a type and a value, which are both atomic. A document may have several different graphs, and graphs can be associated with more than one document; this is not specified in the model.

There are no explicit references. Rather, references are handled implicitly by equivalence classes: if two annotations share the same type and value, they are considered co-referential. To refer to particular documents or other objects, an application or annotator must choose some convention for representing those references as strings and use those as annotation values. This seems problematic: an annotation of type Co-reference Chain and value Chain23 should be equivalent to another of the same type and value, but this is not true for an annotation of type PartOfSpeech and value Noun. Because LDC annotation values are atomic, any representation of complex data structures must define its own reference structure to point into some other representation system.

TIPSTER has a richer formalism, both because of the complexity of the annotation/attribute part of the model and because documents and collections of documents are an explicit part of the model, as are references among all these objects.

The inherent problems with developing a model of a task to be solved in software in isolation from the development of instances of that software are evident in the work of Cassidy and Bird (2000), who discussed the properties of the LDC AG model when stored and indexed in a relational database. At that point the authors added identifier fields to annotations to allow referencing without the equivalent class notion.

Reference Annotation III: GATE GATE version 2 has a reference annotation model that was designed to combine the advantages of the TIPSTER and LDC models:

- Annotation sets are more explicitly graph-based. This feature allows increased efficiency of traversal and simpler editing because offsets are moved from the annotations into a separate node object. In addition, the offsets can be both character and time offsets, thus enabling annotation of multimodal data.

- Multiple annotation sets are allowed on documents. Consider the situation when two people are adding annotations to the same document and later wish to compare and merge their results. TIPSTER would handle this by having an 'annotator' attribute on all the annotations. It is much simpler to have disjoint sets.
- Documents and collections are an essential part of the model, and information can be associated with them in similar fashion to that on annotations.
- All annotations have unique identifiers to allow for referencing.
- An annotation only has two nodes which means that the multiple-span annotations of TIPSTER are no longer supported; the workaround is to store noncontiguous data structures as features of the document and point from there to the multiple annotations that make up the structures.
- The annotation values are extensible (i.e., any classes of object can be added to the model and be associated with annotations).

In addition, both LDC and TIPSTER need an annotation meta-language to describe – for purposes of validation or configuration of viewing and editing tools – the structure and permissible value set of annotations. GATE uses the XML schema language supported by W3C as an annotation meta-language (Cunningham *et al.*, 2002b). These annotation schemas define which attributes and optionally which values are permissible for each type of annotation (e.g., POS, named entity). For instance, a chosen tag set can be specified as permissible values for all POS annotations. This meta-information enables the annotation tools to control the correctness of the user input, thus making it easier to enforce annotation standards.

Data about Language

The preceding sections described language data, information related directly to examples of the human performance of language. This section considers work on data about language or the description of human language competence. Much work in this area has concentrated on formalisms for the representation of the data and has advocated declarative, constraint-based representations (using feature-structure matrices manipulated under unification) as an appropriate vehicle with which "many technical problems in language description and computer manipulation of language can be solved" (Shieber, 1992). One example of an infrastructure project based on Attribute-Value Matrices (AVMs) is ALEP, the Advanced Language Engineering Platform. ALEP aims to provide "the NLP research and engineering community in Europe with an open, versatile, and general-purpose development environment" (Simkins, 1992). ALEP, although open in principle, is primarily an advanced system for developing and manipulating feature structure knowledge bases under unification. It also has several parsing algorithms – algorithms for transfer, synthesis, and generation (Schütz, 1994). As such, it is a system for developing particular types of LRs (e.g., grammars, lexicons) and for doing a particular set of tasks in LE in a particular way.

The system, despite claiming to use a theory-neutral formalism (in fact an HPSG (Head-driven Phrase Structure Grammar)-like formalism), is still committed to a particular approach to linguistic analysis and representation. It is clearly of utility to those in the LE community who use that class of theories and to whom those formalisms are relevant, but it excludes or at least does not support actively those who are not, including an increasing number of researchers committed to statistical and corpus-based approaches.

Other systems that use AVMs include a framework for defining NLP systems based on AVMs (Zajac, 1992); the Eurotra architecture, an 'open and modular' architecture for MT promoting resource reuse (Schütz *et al.*, 1991); the DATR morphological lexicon formalism (Evans and Gazdar, 1996); the Shiraz MT Architecture, a chart and unification-based architecture for MT and (Amtrup, 1999), a unified (Finite State Transducer) FST/AVM formalism for morphological lexicons Zajac (1998a); and the RAGS architecture.

A related issue is that of grammar development in an LE context (see Netter and Pianesi, 1997; Estival *et al.*, 1997). Fischer *et al.* (1996) presented an abstract model of thesauri and terminology maintenance in an OO framework. ARIES is a formalism and development tool for Spanish morphological lexicons (Goni *et al.* 1997).

The Reference Architecture for Generation Systems (RAGS) project (Cahill *et al.*, 1999a,b) has concentrated on describing structures that may be shared among NLG component interfaces. This choice is motivated by the fact that the input to a generator is not a document, but a meaning representation. RAGS describes component I/O using a nested feature matrix representation, but does not describe the types of LR that an NLG system may use or the way in which components may be represented, loaded, and so on. More recently, Mellish *et al.* (2004) presented the RAGS conceptual framework and Mellish and

Evans (2004) discussed the implementation of this framework in several experimental systems and how these systems illustrate a wider range of issues for the construction of SALE for generation.

Indexing and Retrieval

Modern corpora, and annotations upon them, frequently run to many millions of tokens. To enable efficient access to this data, the tokens and annotation structures must be indexed. In the case of raw corpora, this problem equates to information retrieval (IR; also known as document detection), a field with a relatively well-understood set of techniques based on treating documents as bags of stemmed words and retrieving based on relative frequency of these terms in documents and corpora (see van Rijsbergen, 1979). Although these processes are well understood and relatively static, IR is an active research field, partly because existing methods are imperfect and partly because that imperfection becomes more and more troubling in the face of the explosion of web content. There have been several attempts to provide SALE systems in this context.

As noted above, the TIPSTER (1995) program developed a reference model of typical IR component sets. More concretely, this program also developed a communication protocol based on Z39.50 for the detection of interactions between the querying application and search engine (Buckley, 1998). The annotation and attribute data structures described earlier were also applied for IR purposes, although the practical applications of the architecture were found in general to be too slow for the large data sets involved.

GATE (Cunningham *et al.*, 2002a,b) uses an extendable, open-source IR engine, Lucene, to index documents and corpora for full-text retrieval. Lucene also allows indexing and retrieval by custom-provided fields like annotations. The model used to wrap Lucene in GATE is designed for extensibility to other IR systems when required.

Whereas the problem of indexing and retrieving documents is well understood, the problem of indexing complex structures in annotations is more of an open question. The Corpus Query System (Christ, 1994, 1995) is the most-cited source in this area, providing indexing and search of corpora and later of WordNet. Similar ideas have been implemented in CUE (Mason, 1998) for indexing and search of annotated corpora and at the W3-Corpora site (University of Essex, 1999) for searchable on-line annotated corpora. Some work on indexing in the LT XML system was reported in McKelvie and Mikheev (1998). Bird *et al.* (2000a) proposed a query language for the LDC annotation graph model, called AGQL. Cassidy (2002) discussed the use of XQuery as an annotation query language and concluded that it is good for dealing with hierarchical data models like XML, but needs extending with better support for sequential data models, such as annotation graphs.

GATE indexes and retrieves annotations by storing them in a relational database, indexed by type, attributes, and their values. In this way, it is possible to retrieve all documents that contain a given attribute and/or value or to retrieve all annotations of a given type in a corpus, without having to traverse each document separately (Bontcheva *et al.*, 2002; Cunningham *et al.*, 2002b). The query language used is SQL.

Recent Trends and Future Directions

As has become evident from the work reviewed here, there are many tools and architectures, and many of these are focused on subareas of NLP (e.g., dialog speech) or specific formalisms (e.g., HPSG). Each of these infrastructures offers specialized solutions, so it is not likely that there will ever be only one universal architecture or infrastructure. Instead, the focus in recent work has been on 'inter-operability', allowing infrastructures to work together, and reusability, enabling users to reuse and adapt tools with a minimum effort. We review some of these new trends here to see how they are likely to influence the next period of research on SALE.

Toward Multipurpose Repositories

To support the reusability of resources, several repositories have been established; some describe NLP tools (e.g., ACL Natural Language Software Registry), and others distribute language resources, such as corpora and lexicons (e.g., ELRA and LDC). To date, these repositories have remained largely independent of each other, with the exception of such repositories as TRACTOR (Martin, 2001), which contain both corpora in a number of languages and specialized tools for corpus analysis.

As argued in Declerck (2001), there is a need to link the two kinds of repositories to allow corpus researchers to find the tools they need to process corpora and vice versa. The idea is to create a multipurpose infrastructure for the storage and access of both language data and the corresponding processing resources. One of the cornerstones of such an infrastructure are metadata, associated with each resource and pointing at other relevant resources (e.g., tools pointing at the language data that they need and can process). The following section discusses recent research on metadata descriptions for tools and language resources, including handling of multimodal and multilingual data.

Resource Metadata and Annotation Standards

As discussed earlier, there are several reasons why metadata should be part of a component infrastructure (i.e., why it is useful beyond the more narrow scope of providing descriptions of resources in a repository). One dimension that affects the kinds of metadata needed to describe resources is their type: whether they are documents in a corpus, a lexicon, or a tool working on language data. For example, the ISLE Computational Lexicon working group has developed a modular architecture, called MILE, designed to factor out linguistically independent primitive units of lexical information; deal with monolingual, bilingual, and multilingual lexicons; and avoid theoretical bias (Calzolari *et al.*, 2001). Some of these desiderata are relevant also to the problem of resource distribution, as discussed in the section on programmatic access and in Cunningham *et al.* (2000). Multimedia/multimodal language resources (MMLR) pose a different set of problems, and existing standards for tagging textual documents (e.g., XCES; Ide *et al.*, 2000) are not sufficient. Broeder and Wittenburg (2001) provided a metadata vocabulary for MMLR, which encodes information related to the media files (e.g., format and size) and the annotation units used (e.g., POS), as well as the basic information on creator, content, and so on.

Another aspect of improving resource reusability and interoperability is the development of standards for encoding annotation data. Ide and Romary (2002) described a framework for linguistic annotations based on XML and the XML-based RDF and DAML+OIL standards for defining the semantics of the annotations. It provides a link with recent work on formal ontologies and the semantic web and enables the use of the related knowledge management tools to support linguistic annotation. For example, Collier *et al.* (2002) used the popular Protégé ontology editor as a basis for an annotation tool capable of producing RDF(S) annotations of language data in multiple languages.

Open Archives

One of the new research directions is toward 'open archives,' archives aiming to make resources easily discoverable, accessible, and identifiable. This work not only includes language resources, such as corpora and lexicons, but also software tools (i.e., processing resources and development environments). Resource discovery is made possible by metadata associated with each resource and made available in a centralized repository. The recently established Open Language Archives Community (OLAC; Bird and Simons, 2001; Bird *et al.*, 2002) aims to create a worldwide virtual library of language resources through the development of inter-operating repositories and tools for their maintenance and access. OLAC also aims to establish and promote best practices in archiving for language resources. The OLAC infrastructure is based on two initiatives from digital library research: the Open Archieves Initiative and the Dublin Core initiative for resource metadata. Currently, OLAC comprises 12 archives with a cross-archive searching facility.

As argued in Wynne (2002), the current trends toward multilinguality and multimodality suggest that the language resources of the future will span across languages and modalities, will be distributed over many repositories, and will form virtual corpora, supported by a diverse set of linguistic analysis and searching tools. As already discussed, metadata and annotation standards play a very important role here. The other major challenge lies in making existing processing resources accessible over the web and enhancing their reusability and portability.

Component Reusability, Distributed Access, and Execution

To enable virtual corpora and collaborative annotation efforts spanning country boundaries, software infrastructures and tools need to control user access to different documents, types of annotations, and metadata. Ma *et al.* (2002) discussed how this access can be achieved by using a shared relational database as a storage medium, combined with a number of annotation tools based on the annotation graph formalism discussed in the section on the Linguistic Data Consortium. The same approach has been taken in GATE (Cunningham *et al.*, 2002b), in which all LRs and their associated annotations can be stored in Oracle or PostgreSQL. This feature enables users to access remote LRs, index LRs by their annotations, and construct search queries retrieving LRs given annotations or metadata constraints (e.g., find all documents that contain person entities called Bush). User access is controlled at the individual and group level, with read/write access rights specified at LR creation time by their owner (the user who has first stored the LR in the database). Because the storage mechanisms in GATE are separate from the API used for accessing LRs and annotations, the visualization tools and processing resources work on both local and remote data in the same way. Ma *et al.* (2002) discussed a special version of AGTK TableTrans tool created to work with the database annotations. In addition, GATE's database storage model supports other LRs, such as lexicons and ontologies.

The recent development of web services enables integration of different information repositories and services across the Internet and offers a new way of sharing language resources across the Internet. Dalli (2002) discussed an architecture for web-based inter-operable LRs based on SOAP and web services. Work in progress extends this approach to processing resource execution in the context of on-line adaptive information extraction (see Tablan *et al.*, 2003). Both make extensive use of XML for metadata description. However, the benefits of the relational database storage mechanism can still be maintained by providing a conversion layer, which transforms the stored LRs and annotations into the desired XML format when needed. Similarly, Todirascu *et al.* (2002) described an architecture that uses SOAP to provide distributed processing resources as services on the web, both as a protocol for message passing and a mechanism for executing remote modules from the client. Bontcheva *et al.* (2004) reported recent work in upgrading GATE to meet challenges posed by research in semantic web, large-scale digital libraries, and machine learning for language analysis.

Popov *et al.* (2004) presented an application that combines several SALE systems, including GATE and Sesame, to create a platform for semantic annotation called KIM (Knowledge and Information Management). Their paper covered several issues relating to building scaleable ontology-based information extraction.

Measurement

A persistent theme in SALE work has been measurement, quantitative evaluation, and the relationship between engineering practice and scientific theory. To quote Lord Kelvin in a lecture to the Institution of Civil Engineers, in London in 1883.

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the stage of science.

On the other hand, Einstein tells us,

> Not everything that counts can be counted, and not everything that can be counted counts (from a sign hanging in Einstein's office at Princeton University).

Researchers have taken similarly varied approaches to measurement, both of component systems developed using SALE systems and of the success of those systems themselves. The presentation of IBM's TEXTRACT architecture by Neff *et al.* (2004) included an illustration of how the same mechanism can be used for producing both quantitative metrics and for visual feedback to users of the results of automated processing.

Ferrucci and Lally (2004) reported a successor to TEXTRACT called UIMA (Unstructured Information Management Architecture), which is in active development to support the work of several hundred R&D staff working in areas as diverse as question answering and machine translation. The significant commitment of IBM to SALE development indicates the success of the TEXTRACT concept and of architectural support for language processing research.

Prognosis

> The principal defining characteristic of NLE work is its objective: to engineer products that deal with natural language and that satisfy the constraints in which they have to operate. This definition may seem tautologous or a statement of the obvious to an engineer practicing in another, well established area (e.g., mechanical or civil engineering), but is still a useful reminder to practitioners of software engineering, and it becomes near-revolutionary when applied to natural language processing. This is partly because of what, in our opinion, has been the ethos of most Computational Linguistics research. Such research has concentrated on studying natural languages, just as traditional linguistics does, but using computers as a tool to model (and, sometimes, verify or falsify) fragments of linguistic theories deemed of particular interest. This is of course a perfectly respectable and useful scientific endeavor, but does not necessarily (or even often) lead to working systems for the general public (Boguraev *et al.*, 1995).

Working systems for public consumption require qualities of robustness that are unlikely to be achieved at zero cost as part of the normal development of experimental systems in language computation research (Maynard *et al.*, 2002). Investing the time and energy necessary to create robust reusable software is not always the right thing to do, of course; sometimes what is needed is a quick hack to explore some simple idea with as little overhead as possible. To conclude that this is always the case is a rather frequent error, however, and is of particular concern at a time when web-scale challenges to language processing are common.

Also problematic for SALE is the fact that it is not always easy to justify the costs of engineered systems when developers of more informal and short-term solutions have been known to make claims for their power and generality that are, shall we say, somewhat optimistic. The fact that the majority of the language processing field continues to use a SALE system of one type or another indicates that this has been a fruitful pursuit.

Acknowledgments

The authors were partly supported by EPSRC grant GR/N15764/01 (AKT) and by EU grants SEKT, PrestoSpeace, and Knowledge Web.

See also: Human Language Technology; Language Processing: Statistical Methods; Natural Language Processing: System Evaluation; Text Retrieval Conference and Message Understanding Conference.

Bibliography

All websites have been confirmed as live before publication, but may change post-publication.

Amtrup J (1995). *ICE – INTARC Communication Environment user guide and reference manual version 1.4.* University of Hamburg.

Amtrup J (1999). 'Architecture of the Shiraz Machine Translation System.' http://crl.nmsu.edu/shiraz/archi.html.

Artola X, de Ilarraza A D, Ezeiza N, Gojenola K, Hernández G & Soroa A (2002). 'A class library for the integration of NLP tools: definition and implementation of an abstract data type collection for the manipulation of SGML documents in a context of stand-off linguistic annotation.' In *Proceedings of LREC 2002 Third International Conference on Language Resources and Evaluation.* Gran Canaria, Spain. 1650–1657.

Berners–Lee T, Connolly D & Swick R (1999). 'Web architecture: describing and exchanging data. Tech. rep., W3C Consortium.' http://www.w3.org/1999/04/WebData.

Bird S & Liberman M (1999a). *A formal framework for linguistic annotation. Technical report MS-CIS-99-01.* (Philadelphia: University of Pennsylvania. http://xxx.lanl.gov/abs/cs.CL/9903003.

Bird S & Liberman M (1999b). 'Annotation graphs as a framework for multidimensional linguistic data analysis.' In *Towards standards and tools for discourse tagging. Proceedings of the ACL-99 Workshop.* 1–10.

Bird S & Simons G (2001). 'The OLAC metadata set and controlled vocabularies.' In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources.* 27–38.

Bird S, Buneman P & Tan W (2000a). 'Toward a query language for annotation graphs.' In *Proceedings of the Second International Conference on Language Resources and Evaluation.* Athens, Greece.

Bird S, Day D, Garofolo J, Henderson J, Laprun C & Liberman M (2000b). 'ATLAS: a flexible and extensible architecture for linguistic annotation' In *Proceedings of the Second International Conference on Language Resources and Evaluation.*

Bird S, Uszkoreit H & Simons G (2002). 'The open language archives community.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation.*

Boguraev B, Garigliano R & Tait J (1995). 'Editorial.' *Natural Language Engineering 1(1).*

Boitet C & Seligman M (1994). 'The "Whiteboard" architecture: a way to integrate heterogeneous components of NLP systems.' In *Proceedings of COLING '94.* 426–430.

Bontcheva K, Cunningham H, Tablan V, Maynard D & Saggion H (2002). 'Developing reusable and robust language processing components for information systems using GATE.' In *Proceedings of the 3rd International Workshop on Natural Language and Information Systems.* Aix-en-Provence, France: IEEE Computer Society Press.

Bontcheva K, Tablan V, Maynard D & Cunningham H (2004). 'Evolving GATE to meet new challenges in language engineering.' *Natural Language Engineering 10(3/4)*, 349–373.

Booch G (1994). *Object-oriented analysis and design* (2nd edn.). Amsterdam: Benjamin/Cummings.

Bos J, Rupp C, Buschbeck-Wolf B & Dorna M (1998). 'Managing information at linguistic interfaces.' In *Proceedings of the 36th ACL and the 17th COLING (ACL-COLING '98).* 160–166.

Brand S (1994). *How buildings learn.* London: Penguin.

Brew C, McKelvie D, Tobin R, Thompson H & Mikheev A (1999). *The XML Library LT XML version 1.1 User documentation and reference guide.* Edinburgh: Language Technology Group. http://www.ltg.ed.ac.uk.

Brill E (1992). 'A simple rule-based part-of-speech tagger.' In *Proceedings of the Third Conference on Applied Natural Language Processing.*

Broeder D & Wittenburg P (2001). 'Multimedia language resources.' In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources.* 47–51.

Brown P, Cocke J, Pietra S D, Pietra V D, Jelinek F, Lafferty J, Mercer R & Roossin P (1990). 'A statistical approach to machine translation.' *Computational Linguistics 16*, 79–85.

Brugman H, Russel A, Wittenburg P & Piepenbrock R (1998a). 'Corpus-based research using the Internet.' In *Workshop on Distributing and Accessing Linguistic Resources.* Granada, Spain. 8–15. http://www.dcs.shef.ac.uk/~hamish/dalr/.

Brugman H, Russel H & Wittenburg P (1998b). 'An infrastructure for collaboratively building and using multimedia corpora in the humaniora.' In *Proceedings of the ED-MEDIA/ED-TELECOM Conference.*

Buckley C (1998). 'TIPSTER Advanced Query (DN2). TIPSTER program working paper.' (Unpublished).

Burnard L (1995). 'Users reference guide for the British National Corpus.' http://info.ox.ac.uk/bnc.

Busemann S (1999). 'Constraint-based techniques for interfacing software modules.' In *Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP.* Edinburgh: Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Cahill L, Doran C, Evans R, Mellish C, Paiva D, Reape M, Scott D & Tipper N (1999a). 'Towards a reference architecture for natural language generation systems.' *Tech. Rep. ITRI-99-14; HCRC/TR-102.* Edinburgh and Brighton: University of Edinburgh and Information Technology Research Institute.

Cahill L, Doran C, Evans R, Paiva D, Scott D, Mellish C & Reape M (1999b). 'Achieving theory-neutrality in reference architectures for NLP: to what extent is it possible desirable?' In *Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP.* Edinburgh: Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Cahoon B & McKinley K (1996). 'Performance evaluation of a distributed architecture for information retrieval.' In *Proceedings of SIGIR '96.* 110–118.

Calzolari N, Lenci A & Zampolli A (2001). 'International standards for multilingual resource sharing: the isle computational lexicon working group.' In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources.* 39–46.

Carreras X & Padró L (2002). 'A flexible distributed architecture for natural language Analyzers.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation.* 1813–1817.

Cassidy S (2002). 'Xquery as an annotation query language: a use case analysis.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation.*

Cassidy S & Bird S (2000). 'Querying databases of annotated speech.' In *Eleventh Australasian Database Conference.* Canberra: Australian National University.

Cassidy S & Harrington J (2001). 'Multi-level annotation in the Emu speech database management system.' *Speech Communication 33,* 61–77.

Cheong T, Kwang A, Gunawan A, Loo G, Qwun L & Leng S (1994). 'A pragmatic information extraction architecture for the message formatting export (MFE) system.' In *Proceedings of the Second Singapore Conference on Intelligent Systems (SPICIS '94).* B371–B377.

Christ O (1994). 'A modular and flexible architecture for an integrated corpus query system.' In *Proceedings of the Third Conference on Computational Lexicography and Text Research (COMPLEX '94).* http://xxx.lanl.gov/abs/cs.CL/9408005.

Christ O (1995). 'Linking WordNet to a corpus query system.' In *Proceedings of the Conference on Linguistic Databases.*

Clements P & Northrop L (1996). *Software architecture: an executive overview. Tech. Rep. CMU/SEI-96-TR-003.* Pittsburgh: Software Engineering Institute, Carnegie Mellon University.

Collier N, Takeuchi K, Nobata C, Fukumoto J & Ogata N (2002). 'Progress on multilingual named entity annotation guidelines using RDF(s).' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation, Conference.*

Cunningham H (1999). 'A definition and short history of language engineering.' *Journal of Natural Language Engineering 5(1),* 1–16.

Cunningham H (2000). *Software architecture for language engineering.* Ph.D. diss., University of Sheffield. http://gate.ac.uk/sale/thesis/.

Cunningham H (2002). 'GATE, a general architecture for text engineering.' *Computers and the Humanities 36,* 223–254.

Cunningham H & Scott D (2004). 'Introduction to the special issue on software architecture for language engineering.' *Natural Language Engineering 10,* 205–211.

Cunningham H, Freeman M & Black W (1994). 'Software reuse, object-oriented frameworks and natural language processing.' In *New methods in language processing (NeMLaP-1).* Manchester.

Cunningham H, Humphreys K, Gaizauskas R & Wilks Y (1997). 'Software infrastructure for natural language processing.' In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97).* http://xxx.lanl.gov/abs/cs.CL/9702005.

Cunningham H, Peters W, McCauley C, Bontcheva K & Wilks Y (1998). 'A level playing field for language resource evaluation.' In *Workshop on Distributing and Accessing Lexical Resources at Conference on Language Resources Evaluation.*

Cunningham H, Gaizauskas R, Humphreys K & Wilks Y (1999). 'Experience with a language engineering architecture: three years of GATE.' In *Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP.* Edinburgh: Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Cunningham H, Bontcheva K, Peters W & Wilks Y (2000). 'Uniform language resource access and distribution in the context of a General Architecture for Text Engineering (GATE).' In *Proceedings of the Workshop on Ontologies and Language Resources (OntoLex'2000).* Bulgaria: Sozopol. http://gate.ac.uk/sale/ontolex/ontolex.ps.

Cunningham H, Maynard D, Bontcheva K & Tablan V (2002a). 'GATE: a framework and graphical development environment for robust NLP tools and applications.' In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).*

Cunningham H, Maynard D, Bontcheva K, Tablan V & Ursu C (2002b). 'The GATE user guide.' http://gate.ac.uk/.

Dalli A (2002). 'Creation and evaluation of extensible language resources for Maltese.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation.*

Declerck T (2001). 'Introduction: extending NLP tool repositories for the interaction with language data resource repositories.' In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources.* 3–6.

DFKI (1999). 'The Natural Language Software Registry.' http://www.dfki.de/lt/registry/.

EAGLES (1999). *EAGLES recommendations.* http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

Edmondson W & Iles J (1994). 'A non-linear architecture for speech and natural language processing.' In *Proceedings of International Conference on Spoken Language Processing,* vol. 1. 29–32.

Eriksson M (1996). *ALEP.* http://www.sics.se/humle/projects/svensk/platforms.html.

Erman L, Hayes-Roth F, Lesser V & Reddy D (1980). 'The Hearsay II speech understanding system: integrating knowledge to resolve uncertainty.' *Computing Surveys 12.*

Estival D, Lavelli A, Netter K & Pianesi F (eds.) (1997). 'Computational environments for grammar development and linguistic engineering.' Madrid: Association for Computational Linguistics.

Evans R & Gazdar G (1996). 'DATR: a language for lexical knowledge representation.' *Computational Linguistics 22(1).*

Ferrucci D & Lally A (2004). 'UIMA: an architectural approach to unstructured information processing in the corporate research environment.' *Natural Language Engineering 10,* 327–349.

Fikes R & Farquhar A (1999). 'Distributed repositories of highly expressive reusable ontologies.' *IEEE Intelligent Systems 14(2),* 73–79.

Fischer D, Mohr W & Rostek L (1996). 'A modular, object-oriented and generic approach for building terminology maintenance systems.' In *TKE '96: Terminology and Knowledge Engineering.* 245–258.

Gaizauskas R & Wilks Y (1998). 'Information extraction: beyond document retrieval.' *Journal of Documentation 54(1),* 70–105.

Goldfarb C & Prescod P (1998). *The XML handbook.* New York: Prentice Hall.

Goldfarb C F (1990). *The SGML handbook.* Oxford: Oxford University Press.

Goni J, Gonzalez J & Moreno A (1997). 'ARIES: a lexical platform for engineering Spanish processing tools.' *Journal of Natural Language Engineering 3(4),* 317–347.

Görz G, Kessler M, Spilker J & Weber H (1996). 'Research on architectures for integrated speech/language systems in Verbmobil.' In *Proceedings of COLING-96.*

Grishman R (1997). 'TIPSTER architecture design document version 2.3. Tech. rep., DARPA.' http://www.itl.-nist.gov/div894.02/related_projects/tipster/.

Hendler J & Stoffel K (1999). 'Back-end technology for high-performance knowledge representation systems.' *IEEE Intelligent Systems 14(3),* 63–69.

Herzog G, Ndiaye A, Merten S, Kirchmann H, Becker T & Poller P (2004). 'Large-scale software integration for spoken language and multimodal dialog systems.' *Natural Language Engineering 10,* 283–307.

Hobbs J (1993). 'The generic information extraction system.' In *Proceedings of the Fifth Message Understanding Conference (MUC-5).* http://www.itl.nist.gov/div894/894.02/related_projects/tipster/gen_ie.htm.

Ibrahim M & Cummins F (1989). 'TARO: an interactive, object-oriented tool for Building natural language systems.' In *IEEE International Workshop on Tools for Artificial Intelligence.* 108–113.

Ide N (1998). 'Corpus encoding standard: SGML guidelines for encoding linguistic corpora.' In *Proceedings of the First International Language Resources and Evaluation Conference.* 463–470.

Ide N & Romary L (2002). 'Standards for language resources.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation.*

Ide N & Romary L (2004). 'Standards for language resources.' *Natural Language Engineering 10,* 211–227.

Ide N, Bonhomme P & Romary L (2000). 'XCES: an XML-based standard for Linguistic corpora.' In *Proceedings of the Second International Language Resources and Evaluation Conference (LREC).* 825–830.

Jing H & McKeown K (1998). 'Combining multiple, large-scale resources in a reusable lexicon for natural language generation.' In *Proceedings of the 36th ACL and the 17th COLING (ACL-COLING '98).* 607–613.

Kay M, Gawron J & Norvig P (1994). *Verbmobil, a translation system for face-to-face dialog.* Stanford: CSLI.

Koning J, Stefanini M & Deamzeau Y (1995). 'DAI interaction protocols as control strategies in a natural language processing system.' In *Proceedings of IEEE Conference on Systems, Man and Cybernetics.*

Lassila O & Swick R (1999). 'Resource description framework (RDF) model and syntax specification. Tech. Rep. 19990222, W3C Consortium.' http://www.w3.org/-TR/REC-rdf-syntax/.

Lavelli A, Pianesi F, Maci E, Prodanof I, Dini L & Mazzini G (2002). 'SiSSA: an infrastructure for developing NLP applications.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation.*

LREC-1 (1998). *Conference on Language Resources Evaluation (LREC-1).*

LuperFoy S, Loehr D, Duff D, Miller K, Reeder F & Harper L (1998). 'An architecture for dialogue management, context tracking, and pragmatic adaptation in spoken dialogue systems.' In *Proceedings of the 36th ACL and the 17th COLING (ACL-COLING '98)*. 794–801.

Ma X, Lee H, Bird S & Maeda K (2002). 'Models and tools for collaborative annotation.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation.*

Macleod C, Ide N & Grishman R (2002). 'The American National Corpus: standardized resources for American English.' In *Proceedings of the LREC Second International Conference on Language Resources and Evaluation.* 831–836.

Marcus M, Santorini B & Marcinkiewicz M (1993). 'Building a large annotated corpus of English: the Penn Treebank.' *Computational Linguistics 19(2),* 313–330.

Martin W (2001). 'An archive for all of Europe.' In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources.* 11–14.

Mason O (1998). 'The CUE corpus access tool.' In *Workshop on Distributing and Accessing Linguistic Resources.* 20–27. http://www.dcs.shef.ac.uk/~hamish/dalr/.

Maynard D, Tablan V, Cunningham H, Ursu C, Saggion H, Bontcheva K & Wilks Y (2002). 'Architectural elements of language engineering robustness.' *Journal of Natural Language Engineering Special Issue on Robust Methods in Analysis of Natural Language Data 8(2/3),* 257–274.

McClelland J & Rumelhart D (1986). *Parallel distributed processing.* Cambridge, MA: MIT Press.

McKelvie D & Mikheev A (1998). 'Indexing SGML files using LT NSL, IT Index documentation.' http://www.ltg.ed.ac.uk/.

McKelvie D, Brew C & Thompson H (1998). 'Using SGML as a basis for data-intensive natural language processing.' *Computers and the Humanities 31(5),* 367–388.

Mellish C & Evans R (2004). 'Implementation architectures for natural language generation.' *Natural Language Engineering 10,* 261–283.

Mellish C & Scott D (1999). 'Workshop preface.' In *Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP.* Edinburgh: Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Mellish C, Scott D, Cahill L, Evans R, Paiva D & Reape M (2004). 'A reference architecture for generation systems.' *Natural Language Engineering.*

Miller G A (ed.) (1990). 'WordNet: an on-line lexical database.' *International Journal of Lexicography 3(4)* 235–312.

MITRE (2002). 'Galaxy communicator.' http://communicator.sourceforge.net/.

Neff M S, Byrd R J & Boguraev B K (2004). 'The talent system: TEXTRACT architecture and data model.' *Natural Language Engineering.*

Nelson T (1997). 'Embedded markup considered harmful.' In Connolly D (ed.) *XML: principles tools and techniques.* Cambridge, MA: O'Reilly. 129–134.

Netter K & Pianesi F (1997). 'Preface.' In *Proceedings of the Workshop on Computational Environments for Grammar Development and Linguistic Engineering.* iii–v.

Ogden B (1999). 'TIPSTER annotation and the Corpus Encoding Standard.' http://crl.nmsu.edu/Research/Projects/tipster/annotation.

Olson M & Lee B (1997). 'Object databases for SGML document management.' In *IEEE International Conference on Systems Sciences.*

Petasis G, Karkaletsis V, Paliouras G, Androutsopoulos I & Spyropoulos C (2002). 'Ellogon: a new text engineering platform.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation.*

Peters W, Cunningham H, McCauley C, Bontcheva K & Wilks Y (1998). 'Uniform Language resource access and distribution.' In *Workshop on Distributing and Accessing Lexical Resources at Conference on Language Resources Evaluation.*

Poirier H (1999). 'The XeLDA Framework.' http://www.dcs.shef.ac.uk/~hamish/dalr/baslow/xelda.pdf.

Popov B, Kiryakov A, Kirilov A, Manov D, Ognyanoff D & Goranov M (2004). 'KIM – semantic annotation platform.' *Natural Language Engineering.*

Reiter E (1994). 'Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible?' In *Proceedings of the Seventh International Workshop on Natural Language Generation (INLGW-1994).* http://xxx.lanl.gov/abs/CS.cl/9411032.

Reiter E (1999). 'Are reference architectures standardisation tools or descriptive aids?' In *Proceedings of the AISB'99 Workshop on Reference Architectures and Data Standards for NLP.* Edinburgh: Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Reiter E & Dale R (2000). *Building natural language generation systems.* Cambridge: Cambridge University Press.

Rösner D & Kunze M (2002). 'An XML-based document suite.' In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02).*

Schütz J (1994). 'Developing lingware in ALEP.' *ALEP User Group News 1(1).*

Schütz J, Thurmair G & Cencioni R (1991). 'An architecture sketch of Eurotra-II.' In *MT Summit III.* 3–11.

Shieber S (1992). *Constraint-based grammar formalisms.* Cambridge, MA: MIT Press.

Simkins N K (1992). *ALEP user guide.* Luxemburg: cEC.

Simkins N K (1994). 'An open architecture for language engineering.' In *First CEC Language Engineering Convention.*

Sperberg-McQueen C & Burnard L (1994). 'Guidelines for electronic text encoding and interchange (TEI P3). ACH, ACL, ALLC.' http://etext.virginia.edu/TEI.html.

Sperberg-McQueen C & Burnard L (eds.) (2002). *Guidelines for electronic text encoding and interchange (TEI P4)*. TEI Consortium.

Tablan V, Bontcheva K, Maynard D & Cunningham H (2003). 'OLLIE: on-line learning for information extraction.' In *Proceedings of the HLT-NAACL Workshop on Software Engineering and Architecture of Language Technology Systems*.

Tablan V, Ursu C, Bontcheva K, Cunningham H, Maynard D, Hamza O, McEnery T, Baker P & Leisher M (2002). 'A Unicode-based environment for creation and use of language resources.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation*.

Object Management Group (1992). *The common object request broker: architecture and specification*. New York: John Wiley.

TIPSTER (1995). 'The generic document detection system.' http://www.itl.nist.gov/div894/894.02/related_projects/tipster/gen_ir.htm.

Todirascu A, Kow E & Romary L (2002). 'Towards reusable nlp components.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation*.

Tracz W (1995). 'Domain-specific software architecture (DSSA) frequently asked questions (FAQ).' http://www.oswego.com/dssa/faq/faq.html.

University of Essex (1999). 'Description of the W3-Corpora web-site.' http://clwww.essex.ac.uk/w3c/.

van Rijsbergen C (1979). *Information retrieval*. London: Butterworths.

Veronis J & Ide N (1996). 'Considerations for the reusability of linguistic software. Tech. rep., EAGLES.' http://w3.lpl.univ-aix.fr/projects/multext/LSD/LSD1.html.

von Hahn W (1994). 'The architecture problem in natural language processing.' *Prague Bulletin of Mathematical Linguistics 61*, 48–69.

Wolinski F, Vichot F & Gremont O (1998). 'Producing NLP-based on-line contentware.' In *Natural Language and Industrial Applications*. http://xxx.lanl.gov/abs/cs.CL/9809021.

Wynne M (2002). 'The language resource archive of the 21st century.' In *Proceedings of the LREC 2002 Third International Conference on Language Resources and Evaluation*.

Young S, Kershaw D, Odell J, Ollason D, Valtchev V & Woodland P (1999). *The HTK book (Version 2.2)*. Cambridge: Entropic Ltd. ftp://ftp.entropic.com/pub/htk/.

Yourdon E (1989). *Modern structured analysis*. New York: Prentice-Hall.

Zajac R (1992). 'Towards computer-aided linguistic engineering.' In *Proceedings of COLING '92*. 828–834.

Zajac R (1997). 'An open distributed architecture for reuse and integration of heterogenous NLP components.' In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*.

Zajac R (1998a). 'Feature structures, unification and finite-state transducers.' In *International Workshop on Finite State Methods in Natural Language Processing*.

Zajac R (1998b). 'Reuse and integration of NLP components in the Calypso architecture.' In *Workshop on Distributing and Accessing Linguistic Resources*. 34–40. http://www.dcs.shef.ac.uk/~hamish/dalr/.

Relevant Websites

http://www.tc37sc4.org – ISO standardization.
http://www.ldc.upenn.edu – Linguistic Data Consortium.
http://www.info.ox.ac.uk – British National Corpus.
http://www.openarchives.org – Open Archives Initiative.
http://www.dublincore.org – Dublin Core Initiative for Resource Metadata.
http://www.openrdf.org – Knowledge and information management.