# Uniform Language Resource Access and Distribution

**Hamish Cunningham, Wim Peters, Clare McCauley, Kalina Bontcheva, Yorick Wilks**

Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield, S1 4DP, UK
{hamish, wim, clarem, kalina, yorick}@dcs.shef.ac.uk

Abstract

The reuseability and accessibility of lexical and linguistic resources often requires substantial programming overhead and detailed knowledge of the structure and nature of resource-specific information. Each resource has its own representation syntax and covers a particular subset of linguistic phenomena. This paper will discuss ways to overcome these barriers to resource reuse and presents a new method for distributing and accessing language resources within the GATE environment (General Architecture for Text Engineering). In this environment linguistic resources are modelled in an object-oriented way, yielding an initial set of resource-specific inheritance hierarchies. The exploitation of commonalities between resources leads to an incremental integration of their hierarchies into one common object model (whilst preserving the originals). The representation of objects within this model will be geared towards generally accepted linguistic concepts, which will be realised in a standardised format. This approach has several advantages: the linguistic knowledge is modelled in a conceptual and maximally uniform way; the linguistic resources are represented by maximally uniform data structures; overlaps and differences between resources can be identified by means of the common object model; the model provides a level playing field for the evaluation of resources. In describing parts of the model we will stick closely to the terminology and structures of the existing resources. This is not a new standards initiative, but a way to build more effectively on previous initiatives. The distribution mechanism for the object model will use WWW protocols and enable remote usage of resources without the necessity of local installation.

## 1. Introduction and Background

In general, the reuse of NLP data resources (such as lexicons or corpora) has exceeded that of algorithmic resources (such as lemmatisers or parsers) (Cunningham *et al.*, 1994). Although data for specific tasks often needs to be created from scratch or extracted from existing resources, nevertheless, for English at least, there exists a sizeable number of resources that are reusable and whose contents can serve multiple purposes. However, there are still two barriers to data resource reuse:

1. each resource has its own representation syntax and corresponding programmatic access mode (e.g. SQL for Celex, C or Prolog for WordNet);
2. resources must generally be installed locally to be usable, and how this is done depends on what operating systems are supported etc., which varies from site to site.

A consequence of (1) 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 organised 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 which brings significant overhead. Work which seeks to investigate or exploit commonalities between resources has first to build a layer of access routines on top of each resource. So, for example, if we wish to do task-based evaluation of lexicons, by measuring the relative performance of an information extraction system with different instantiations of lexical resource, we might have to write code to translate several different resources into SQL or some other common format.

A consequence of (2) is that there is no way to "try before you buy": no way to examine a data resource for its suitability for one's needs before licencing it. Correspondingly there is no way for a resource provider to give limited access to their products for advertising purposes, or gain revenue through piecemeal supply of sections of a resource.

This paper will discuss ways to overcome these barriers. We will present a new method for distributing and accessing language resources involving the development of a common programmatic model of the various resources types, implemented in UML (Unified Modelling Language, the new lingua franca of object-oriented modelling (Fowler, 1997)) and Java, along with a distributed server for non-local access (Zajac, 1997). This model and server are being designed to integrate with GATE (General Architecture for Text Engineering (Cunningham *et al.*, 1997)) and go under the provisional title of an Active CREOLE Server. (CREOLE: Collection of REusable Objects for Language Engineering). Currently, CREOLE includes only algorithmic objects, but will be extended to data objects.

The issues of standards is a vexed one: experience with repositories of lexical materials (e.g. the CRL Consortium for Lexical Research 1989-93) suggested that if resources had to have standardised formats, they would not be deposited or used. The success of WordNet worldwide is a demonstration of how researcher choice can defy any committee's standards. What we propose here is quite different from projects like SEAL (Evans & Kilgariff, 1995) that attempt to conflate different lexical resources: in what we propose, the resources retain their integrity, or "native" structure. We propose, via an object oriented methodology, a standardised taxonomy and structure only as an index to the links between lexical objects with the same function in the various resources. Those wishing to research the merging of resources will be one of the beneficiaries of our work, but this is not our primary aim.

## 2. Objectives

Our aim is to create a common object-oriented model for language data resources that encapsulates the union of the linguistic information contained in a range of resources, and will encompass as many object hierarchies as there are resources. At the top of the resource hierarchies are very general abstractions (e.g. the 'head concepts' in any thesaurus like WordNet); 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. Generalisations will be made over different object types in the resources, and the object hierarchies will be merged at whatever levels of description are appropriate. No single view of the data will be imposed on the user, who may chose to stay with the "original" representation of a particular resource, or to access a model of the commonalities between several resources, or a combination of both: our aim is above all inclusive, and this is *not* a new standards initiative, but a way to build on previous initiatives ("standardised" and otherwise).

The object-oriented architecture will represent the linguistic data in a conceptual way. The user will be able to think about a word and its properties without having to assess the actual structure of the data storage. This is in contrast to other data structures such as the relational format where the information about objects is often scattered over many relations or records (Elmasri & Navathe, 1994) (p.665), which obscures the conceptual transparency of the database. It is also more efficient than textual representation in e.g. SGML, which only handles trees fluently, and does not support random access.

A common object model, sitting on top of the resource-specific models, will allow a uniform access procedure for all resources. This will reduce the amount of programmatic overhead referred to above, and be beneficial to the user in other ways. The object-oriented model will allow the user to:

- select the information needed in a conceptual manner, without first having to select the most appropriate resources by detailed examination and comparison, or to extract the information needed in a way determined only by the original structure of the resource in question;
- identify overlaps and differences between the resources;
- assess the usability of each resource for specific language processing tasks.

The users of linguistic data can be divided roughly into two camps: linguists and computer scientists. Conceptual unification of linguistic data structures will help both in different ways. The former will no longer be confused by having to reassess each resource and dedicate time to computational tasks which should in principle be unnecessary for their work. The latter are often lacking linguistic knowledge, and will benefit because they will only have to interpret one kind of data structure.

In general, an object-oriented design will enhance the conceptualisation, reusability and integratability of the resources for all types of users.

To summarise, we are developing three things:

1. an OO model of each resource, documented in UML and accessible in Java, that sticks very closely to the structures and terminology of the resource as it currently exists (e.g. for WordNet, accessor methods on our objects will mirror the functionality of the existing C and Prolog APIs);
2. a unifying model layered on top of the resource-specific models, that captures generalisations about them (implemented in the same way);
3. a data management substrate that handles distributed storage and efficient access to the data underlying the models of 1 and 2.

We will discuss 1 and 2 further below: for more details of 3 see (Zajac, 1997) or `http://www.dcs.shef.ac.uk/research/groups/alp/gate`.

## 3. Resources

We have begun work on covering the following resources: WordNet (Miller (Ed.), 1990), Comlex (Macleod & Grishman, 1994), Celex (Burnage, 1990), EuroWordNet (Vossen *et al.*, 1997), CRL-LDB (Wilks *et al.*, 1996) and Mikrokosmos (Onyshkevych *et al.*, 1996). They are available from the owners or distribution organisations such as ELRA and the Linguistic Data Consortium. We will not in general seek to redistribute them, but to provide the code that maps them into our object model in such a way as to enable others in the community to use their own separately licenced copies with the model and server.

These resources vary in both linguistic scope and granularity, and each resource covers one or more linguistic areas. Comlex contains mostly orthographic and morphosyntactic information; Celex mainly has orthographic, morphological, syntactic and phonological data, whereas WordNet, EuroWordNet and MikroKosmos provide mostly semantic information. They therefore constitute a representative sample of lexical data.

## 4. Building the object models

As we have said previously, our approach is to both:

1. model the native structure of the resources separately and
2. integrate the resources at various levels of description

For task 1, we model exactly the data and the conventions (or terminology) in each resource. The modelling is guided by the existing data structures and whatever documentation is available. For instance, Celex provides several types of phonetic transcription for lemmata and wordforms in a relational structure. These types result in different phonetic transcription objects: plain, syllabified and syllabified with stress. Since we are modelling only resource-specific choices at this stage, we maintain all of this information at this level of granularity and the conventions that are used to encode it. We follow the same procedure for each resource.

For task 2, we compare, extend and incrementally combine the resource-specific models, thus creating a normalisation of resource-specific information into one common object model. However, by continuing to maintain the base, "pure", models of the individual resources, the user always has the option to use the resource-specific format they are familiar with.

Pooling these resources together into one resource can yield a richer common data structure than any single resource can provide. Our common model will maximise access to the total range and variety of data whilst maintaining the individual distinctions between each resource and combining them in a linguistically-controlled environment.

The integration of the resource-specific models is, however, a complicated process. Firstly, when creating object models we need to stay true to the specific properties of the resources, but also want to guarantee as much as possible that the resources can be compared and combined where they share data or complement each other. Secondly, as table 1 shows via a comparison of a number of linguistic features from the four resources, not all resources have the same coverage or granularity of description and none of them share the same linguistic conventions to describe (potentially) the same features.

The task we are thus faced with is multi-faceted: not only *how* and *where* to integrate - but which conventions and descriptive apparatus to use in the integration in order to pull commonalities together into one standardised format.

With respect to the first difficulty, one approach is to integrate where the resources show obvious corresponding representations of linguistic information, e.g. where they share the same name for a particular attribute (and cover the same linguistic ground by means of that attribute name and the corresponding values) these correspondences are straightforward. For example, the attributes 'lemma' in Celex and 'word' in WordNet can easily be classified as subclasses of a postulated common class 'lemma' in the merging process.

The complicating factor is that each resource often has its own specific data structure with unique attributes and value sets that it does not share with other resources. Whilst the resource-specific models will mirror these idiosyncrasies in task 1, depending on the degree of their compatibility, we may have to tune these choices in task 2 to allow integration. This tuning process can work both at a broad level of linguistic description (for instance, at the level of morphology) and at a finer-grained level of description (for instance, at the level of subcategorisation).

Integration at a broad or high-level has an advantage over our first approach in that it deals with all phenomena rather than limiting the integration to shared likenesses. It is not without its own difficulties however. In order for us to link resource-specific information to a superordinate object SUBCATEGORISATION, for example, we may have to 'unpack' the data if it is combined with other information in the resource. One such case is CRL-LDB which combines morphosyntax and subcategorisation in one grammar code, for example:

```
   [C3] count noun followed by infinitive
with 'to'
```

Whilst it is important to mirror this idiosyncrasy in task 1, it seems spurious to maintain it for task 2.

Once these problems are overcome, we will have a common model where each resource-specific representation is initially classified as a separate subcategorisation object linked to a superclass SUBCATEGORISATION. We then face a further difficulty: whilst the idiosyncrasies of each resource's representation are preserved and the resources are mapped at a high level of linguistic abstraction, their interrelationships at more specific levels of description are not clear because of the lack of uniformity. However we may only link at a more specific level once the differences in resource granularity and coverage and in the descriptive apparatus and conventions used to annotate this information are factored out.

A simple case study of the encoding of subcategorisation information for the verb 'build' in four different resources shows how this might be achieved. As it can be seen in Figure 1, CRL-LDB uses idiosyncratic codes (such as `T1` for transitive verbs) for syntactic subcategorisation patterns and these codes are themselves explained by means of a sentence definition. Comlex also uses conventional linguistic labels but exemplifies patterns by means of a formal grammatical representation (e.g. `NP-PP`). WordNet classifies subcategorisation with surface strings containing explicit semantic preference (e.g. `Somebody Vs Something`). Celex distinguishes verbal subcategorisation by typing verbs with 'transitive', 'intransitive' and 'ditransitive'.

In order for us to link at this specific level, we need a common representation into which the data of the resources can be moulded. For this purpose we need to choose a canonical representation format. We can choose between:

1. using one resource's conventions as the descriptive model of all the resources
2. creating a new standard representation, or
3. using existing standards for linguistic encoding such as EAGLES (ILC-CNR, 1996).

The best option may be to adhere to existing standards for representational purposes. The EAGLES slot and filler system is already being applied in PAROLE which uses a derivative of the standard in its modelling of lexical resources via the object-oriented model GENELEX. However, although this is a general model which can be applied to any type and degree of syntactic information, there are some language specific difficulties that may require the standards to be refined. Moreover, unlike the rest of the options, this

| | |
|---|---|
| **CELEX** | transitive |
| | intransitive |
| | ditransitive |
| **COMLEX** | (PP :PVAL (into, for, over, on, upon)) |
| | (NP-PP :PVAL (around, from, of, upon, out of, on, into, up)) |
| | (PART-PP :ADVAL (out, up) |
| | :PVAL (along, over, onto, to)) |
| | (NP-P-NP-ING :PVAL (on, upon)) |
| | (PART-NP :ADVAL (up)) |
| | (NP-FOR-NP) |
| | (NP) |
| | (NP-AS-NP)) |
| **CRL-LDB** | [T1] transitive with one object followed by one or more nouns or pronouns |
| **WordNet** | Somebody —-s something |
| | Somebody —-s |
| | Something —-s |
| | Something —-s something |

Figure 1: Subcategorisation information for the verb 'build'

| | CRL-LDB | CELEX | COMLEX | WordNet |
|---|---|---|---|---|
| **Orthography:** | | | | |
| spelling variants | Y | Y | Y | Y |
| syllabification | Y | Y | | |
| **Morphosyntax:** | | | | |
| part of speech | Y | Y | Y | Y |
| grammatical subcategory | Y | Y | Y | |
| inflectional characteristics | Y | Y | Y | Y |
| **Morphology:** | | | | |
| derivational information | | Y | | |
| **Phonology:** | | | | |
| pronunciation | Y | Y | | |
| **Syntax:** | | | | |
| verb subcategorisation | Y | Y | Y | Y |
| **Semantics:** | | | | |
| sense distinction | Y | | | Y |
| verb argument preferences | Y | | | Y |
| ontological classification | Y | | Y | Y |
| semantic relations | | | | Y |
| domain | Y | | | |

Table 1: Linguistic features contained in four resources

is the most radical in that the result of the modelling may be a structure unlike anything that is actually in any of the resources (e.g. a transitive verb encoded `T1` will be decomposed into a number of position slots [SVO] each of which can have a number of realisations (subject can be realised by NP, Clause etc.). As a result of this encoding, the data will look very different from what is actually there in the resource - even though it is encoding the same information. It is still debatable whether we need or want to decompose such phenomena as subcategorisation to this degree of explicitness.

## 5. Conclusion

The initial development of this object hierarchy will not yield anything substantively new, but will improve access to existing resources and aid practical exploitation of existing standards. This material will be available in distributed form and targeting multiple database backends.

An important benefit of this work will be the creation of an appropriate environment for the evaluation of resources. Whereas at present it is practically very difficult to treat different resources in the same way for purposes of evaluation, the new access method we propose will make this much more feasible, and provide a firm basis for comprehensive evaluation efforts.

Our aims are reuse, inclusiveness and flexibility and the results of this work should be adaptable for any language processing task. The data required can be chosen at any level of granularity or resource-specificity.

Of course, the production of a common model that fully expresses all the subtleties of all available resources would be a large undertaking, but we believe that it can be done incrementally, with useful results at each stage. We propose to stop decomposing the object structure of resources at a fairly high level, leaving the developer to handle the original data structures of the resources at the leaves of the forest. Even at this stage we still expect substantial benefit from uniform access to higher level structures.

## 6. References

Burnage, G. 1990. *CELEX, A Guide for Users*. CELEX Centre for Lexical Information, Nijmegen, The Netherlands.

Cunningham, H., Freeman, M., & Black, W.J. 1994. Software Reuse, Object-Oriented Frameworks and Natural Language Processing. *In: Proceedings of the conference on New Methods in Natural Language Processing (NeMLaP-1)*.

Cunningham, H., Humphreys, K., Gaizauskas, R., & Wilks, Y. 1997 (Mar.). Software Infrastructure for Natural Language Processing. *Pages 237–244 of: Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*. Available as `http://xxx.lanl.gov/ps/9702005`.

Elmasri, R., & Navathe, S.B. 1994. *Fundamentals of Database Systems*. Benjamin/Cummings, Redwood City, California.

Evans, R., & Kilgariff, A. 1995. MRDs, Dictionaries, and How to Do Lexical Engineering. *Pages 125–132 of: Proceedings of the 2nd language engineering convention*. London, UK.

Fowler, M. 1997. *UML Distilled: Applying the Standard Modelling Language*. Addison-Wesley.

ILC-CNR. 1996. *Preliminary Recommendations on Subcategorisation EAGLES document EAG-CLWG -SYNLEX/P*. Tech. rept. Available from `http://www.ilc.pi.cnr.it/EAGLES/browse.html`.

Macleod, C., & Grishman, R. 1994. *COMLEX Syntax Reference Manual*. Proteus Project, NYU.

Miller (Ed.), G. A. 1990. WordNet: An on-line Lexical Database. *International Journal of Lexicography*, **3**(4), 235–312.

Onyshkevych, B., Boyan, & Nirenburg, S. 1996. Microkosmos. *Machine Translation 10:1-2 (Special Issue on building lexicons for MT)*.

Vossen, P., Diez-Orzas, P., & Peters, W. 1997. The Multilingual Design of EuroWordNet. *In: Proceedings of the ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Wilks, Y., Guthrie, L., & Slator, B. 1996. *Electric Words*. MIT Press, Cambridge, MA.

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)*.