Collaborative Resource Development and Delivery

Workshop Programme

09:00 –09:15 – Welcome and Overview

09:15 –10:00 – Discussion paper

The MASC/MultiMASC Community Collaboration Project: Why You Should Be Involved and How
Nancy Ide, Collin Baker, Christiane Fellbaum, Rebecca Passonneau

10:00 –10:30 – Discussion: Strategies to Engage the Community in Collaborative Annotation

10:30 –11:00 Coffee break

11:00 –11:30 – Invited talk

Towards a Linguistic Linked Open Data Cloud
Christian Chiarcos

11:30 –12:30 – Collaborative annotation task and discussion

12:30 –14:00 – Lunch break

14:00 –15:30 Paper session

Annotated Corpora in the Cloud: Free Storage and Free Delivery
Graham Wilcock

Guidance through the Standards Jungle for Linguistic Resources
Maik Stührenberg, Antonina Werthmann, Andreas Witt

Supporting Collaborative Improvement of Resources in the Khresmoi Health Information System
Lorraine Goeuriot, Allan Hanbury, Gareth J. F. Jones, Liadh Kelly, Sascha Kriewel, Ivan Martinez Rodriguez, Henning Müller, Miguel A. Tinte

15:30 –16:00 – Demonstrations

16:00 –16:30 Coffee break

16:30 –17:40 – Paper session

Building Parallel Corpora Through Social Network Gaming
Nathan David Green

Three Steps for Creating High-Quality Ontology Lexica
John McCrae, Philipp Cimiano

PromONTotion: Creating an Advertisement Thesaurus By Semantically Annotating Ad Videos Through Collaborative Gaming
Katia Lida Kermanidis, Emmanouil Maragkoudakis

The Phrase Detective Multilingual Corpus, Release 0.1
Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, Luca Ducceschi

17:40 –18:00 – Closing

Editors

Nancy Ide	Vassar College, USA
Collin Baker	ICSI/UC Berkeley, USA
Christiane Fellbaum	Princeton Univeristy, USA
Rebecca Passonneau	Columbia University, USA

Workshop Organizers/Organizing Committee

Nancy Ide	Vassar College, USA
Collin Baker	ICSI/UC Berkeley, USA
Christiane Fellbaum	Princeton Univeristy, USA
Rebecca Passonneau	Columbia University, USA

Workshop Programme Committee

Nicoletta Calzolari	ILC/CNR, Italy
Bob Carpenter	Alias I, Inc., USA
Chris Cieri	LDC, University of Pennsylvania, USA
Bill Dolan	Microsoft Corp., USA
Dan Flickinger	Stanford University, USA
Terry Langendoen	NSF and University of Arizona, USA
Massimo Poesio	University of Trento, Italy
Sameer Pradhan	BBN Technologies, USA
James Pustejovsky	Brandeis University, USA
Owen Rambow	Columbia University, USA
Manfred Stede	Universität Potsdam, Germany

Table of contents

Invited Talk

Author Index

Introduction

A confluence of needs and activities points to a new emphasis in computational linguistics to address lexical, propositional, and discourse semantics through corpora. A few examples are:

- the demand for high quality linguistic annotations of corpora representing a wide range of phenomena, especially at the semantic level, to support machine learning and computational linguistics research in general;
- the demand for high quality annotated corpora representing a broad range of genres that are flexible and extensible as need demands;
- the demand for high quality lexical and semantic resources to incorporate into the annotation process, and for the annotation process to produce;
- the need for easy-to-use, open access to all of these resources for everyone.

Such resources can be very costly to produce, due to the need for manual creation or validation to ensure quality. Therefore, to answer the growing need and lower the costs of resource creation and enhancement, there is a movement within the community toward collaborative resource development, including collaborative corpus annotation and collective creation/enhancement of lexical resources and knowledge bases. Collaborative development encompasses both engaging the community in annotation and development of common resources, as well as crowdsourcing, gaming, and similar solutions. The papers in this workshop address both of these approaches to collaborative development, as well as software to support this development and other issues such as the role of standards for collaboratively created resources.

This workshop was motivated by a meeting of eminent researchers and developers of language resources, held at Columbia University in New York City in October, 2011. The goal of the meeting was to explore ways to involve the natural language processing community in the development of language resources—most notably, annotated linguistic corpora—in order to offset the high costs of resource creation. The focus of the discussions was the Manually Annotated Sub-Corpus (MASC) (http://www.anc.org/MASC), a community-based collaborative annotation project that is intended to provide the basis for development of a resource that is richly annotated for both variety and variants of linguistic phenomena. Among the conclusions of that workshop was a decision to broaden the discussion to include the community a whole by holding a workshop at LREC. This workshop therefore includes a special session devoted to strategies for engaging the community in collaborative linguistic annotation projects such as MASC.

The workshop also includes a collaborative annotation task that will engage all participants in the annotation of multiple phenomena over a common text. A following discussion session considers the results in order to address issues such as the level of agreement among the participants on the various tasks, and what it suggests in terms of the viability of collaborative annotation and crowdsourcing for creating high-quality linguistic annotations; and ways in which annotations on multiple levels may be used collectively to improve overall quality and contribute to analysis.

Annotated Corpora in the Cloud: Free Storage and Free Delivery

Graham Wilcock

University of Helsinki
graham.wilcock@helsinki.fi

Abstract

The paper describes a technical strategy for implementing natural language processing applications in the cloud. Annotated corpora can be stored in the cloud and queried in normal web browsers via user interfaces implemented in the described framework. A key aim of the strategy is to exploit the free storage and processing that is available in the cloud, while avoiding lock-in to proprietary infrastructure. A half-million-word annotated corpus application is described as a working example of the strategy.

1. Introduction

The paper describes a technical strategy for designing and implementing natural language processing applications in the cloud in such a way that annotated corpora can be queried and displayed in ordinary web browsers. There are many different strategies for cloud computing, but rather than giving a superficial review of a variety of alternatives, the paper focusses on describing one specific approach. This approach can be summarized as "open source front-end, proprietary (but free) back-end".

The paper focusses exclusively on approaches that offer free storage of the corpora in the cloud, and free delivery of the corpora contents and annotations to the web browser. The example corpus application that demonstrates these approaches does not currently support collaborative development of the annotations.

Like many other applications, corpus applications can be regarded as having three main parts. The "front-end" is the user interface, typically consisting of a set of web pages and ways to navigate between them. The "back-end" is where the data is stored, typically in a database. The application processing takes place somewhere "in the middle".

This division into three parts is well-known in computer science as the "model, view, controller" design pattern. Here, the back-end database is the model, the front-end user interface is the view, and the application processing in the middle is the controller.

In the case of a cloud computing application, the data is stored in some special kind of cloud data store and the processing is done in a special cloud run-time environment, but it is important that the user interface works in an ordinary web browser.

The component parts of the technical strategy are described in the next section. Section 3. then reviews related work. Section 4. describes an implemented example application, in which an annotated corpus is stored in the cloud and is queried from ordinary web browsers. Problems and solutions from this implementation are discussed in Section 5., and Section 6. presents conclusions.

2. A Technical Strategy

This section sets out a technical strategy for design and implementation of cloud-based applications. A key aim of the strategy is to take advantage of the free storage and processing quotas that are available in the cloud, while avoiding lock-in to one specific proprietary infrastructure. We believe that this can be achieved by appropriate choices of the front-end and back-end components.

The choices proposed in this technical strategy are Django, an open source web framework, and Google App Engine, a proprietary cloud computing platform. The strategy of "open source front-end, proprietary (but free) back-end" is therefore more specifically implemented as "Django front-end, Google App Engine back-end".

2.1. The cloud computing framework

Google App Engine (`http://code.google.com/appengine`) is a platform for running web apps in the cloud on Google's infrastructure. One of the motivations for choosing App Engine as the preferred cloud computing framework is that Google currently allow applications to be run entirely free of charge, as long as they stay within certain quotas. The quotas apply to several dimensions: processing power, overall storage capacity, individual file sizes, response times. Significant applications can be implemented within the free quotas, and can be hosted on Google's infrastructure with zero running costs.

Like other cloud frameworks, there are no maintenance costs for server hardware or server software. The Google architectures are massively scalable. If the quotas are exceeded App Engine is no longer free, but this will only occur if the applications are massively successful, which is a very desirable "problem".

Even in this case, there is no obligation to pay for the additional resources required to meet the higher demand. The application can simply be restricted to the free quotas. The users will experience this as longer response times or reduced service availability at times of high demand, but there will be no charges unless billing has been authorized.

When selecting a framework that is currently free of charge, the danger of lock-in to the specific technology must be considered, in case charging is introduced at some time in the future. This important question is addressed in Section 5.3..

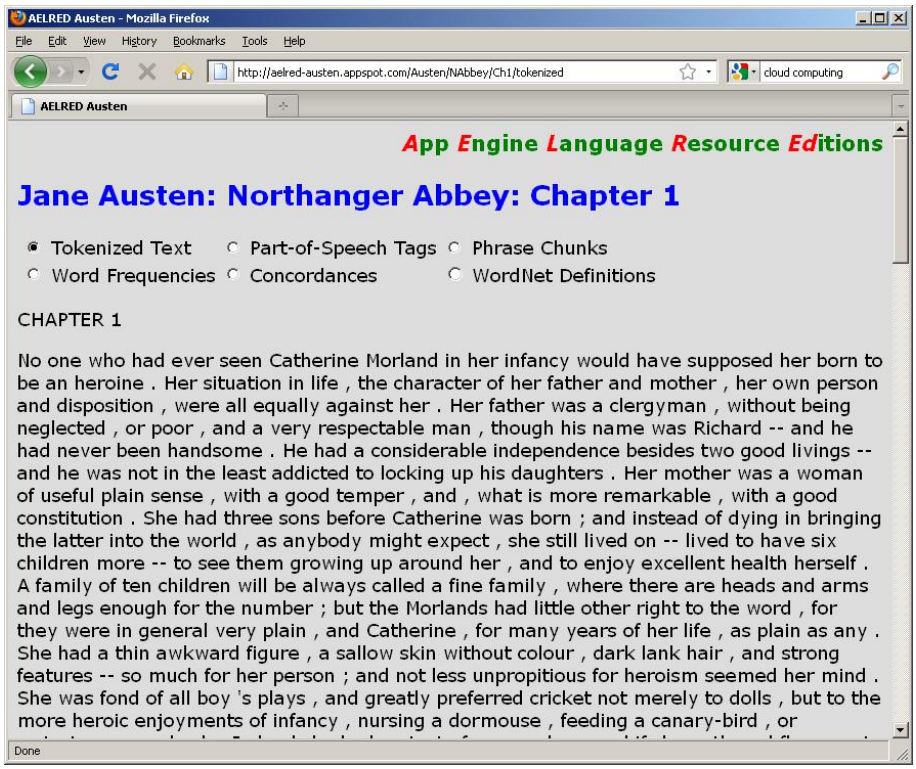

Figure 1: Example application: tokenized text.

2.2. The web app front-end

App Engine includes its own simple web app framework, but other standards-compliant front-end frameworks can be imported. Our strategy uses Django (`http://www.djangoproject.com`), a successful and widely-used open source Python web app framework (Holovaty and Kaplan-Moss, 2009).

Django provides a wide range of components that speed up web app development. One of the most important is the Django template engine, which supports dynamic generation of HTML web pages. The template slots are filled-in with the relevant information from the specific context, using appropriate filters, conditionals and loops.

Collections of templates can be managed by organizing them into template hierarchies, where more specific templates inherit information from base templates. Inheritance can take place at several different levels.

Django also provides a clean way to manage the mapping between the application URLs and the processing code that handles the HTTP requests, and an object-relational mapping (ORM) between the object-oriented Python processing code and the back-end relational database models.

2.3. The database back-end

Django is normally used with an SQL database. This can be a full-scale database system such as MySQL or a light database such as SQLite3. By contrast, App Engine is normally used with its own non-relational datastore, which is based on Google's BigTable technology.

The advantages of using the App Engine datastore are that its use is free within the quotas, while being massively scalable if required. However, there are two main disadvantages. First, the non-relational "NoSQL" architecture is less familiar to most developers than standard SQL databases. Second, there could be a danger of lock-in to Google's proprietary technology.

The example application described in Section 4. originally used the App Engine datastore back-end together with the App Engine web app front-end. This version can be seen at `http://aelred-austen.appspot.com`. The prototype has subsequently been re-implemented to make it portable, so that either a MySQL relational database or an App Engine non-relational datastore can be used.

It is possible to combine a Django front-end with an App Engine datastore back-end. This version of our example application can be seen at `http://django-appeng.appspot.com`.

It has recently become possible to use a MySQL database with App Engine in the Google Cloud SQL service (`http://code.google.com/p/googlecloudsql`). Another version of our example application, combining Django and MySQL with App Engine, can be seen at `http://django-mysql.appspot.com`.

2.4. Application processing

In our strategy the application processing that connects the front-end user interface and the back-end database is written in Python. We use NLTK Natural Language Toolkit (Bird et al., 2009) for the language processing tasks, where possible, while organizing the user interaction within the Django framework.

NLTK (`http://www.nltk.org`) provides a set of tools and resources for natural language processing. Like Django, NLTK is a successful and widely-used open source

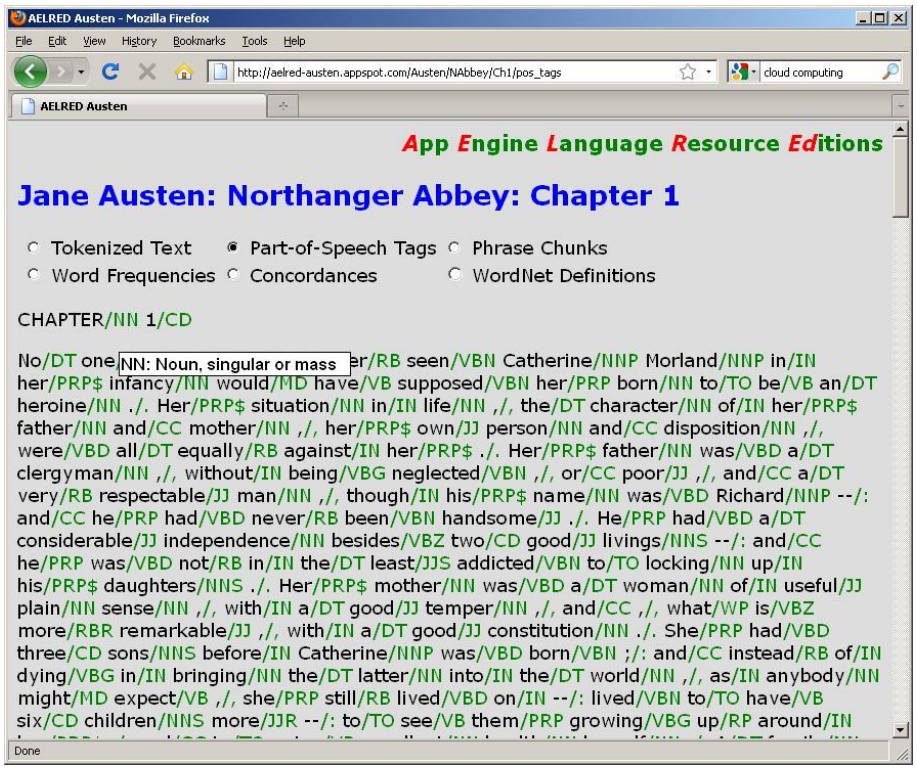

Figure 2: Example application: part-of-speech tags and a tooltip explanation.

Python toolkit.

The ready-made NLTK tools include a sentence boundary detector `nltk.sent_tokenize()`, a word tokenizer `nltk.word_tokenize()`, a part-of-speech tagger `nltk.pos_tag()` and a classifier-based named entity recognizer `nltk.ne_chunker()`.

In addition, NLTK includes useful wordlists, such as lists of stopwords. NLTK also includes a complete version of WordNet, and a convenient Python-WordNet interface.

However, there are some technical issues in using these tools with Google App Engine, which are discussed further in Section 5.2..

2.5. Annotation format

The most widely-used markup language for linguistic annotation of texts is XML. While it is generally agreed that XML should be used for external interchange of linguistic annotations, as it is the global standard for data interchange, it is not necessarily the best choice for internal representation of annotations.

When working in Python it is more convenient to use JSON as an internal representation. Python objects can be serialized easily and quickly to JSON strings, and JSON strings can be deserialized easily and quickly to Python objects. Our strategy therefore recommends storing linguistic annotations in JSON format in the back-end database. Typically, complete chapters of novels can be stored as long text strings in the database, even when expanded by adding linguistic annotations.

3. Related Work

Corpus linguistics is usually done with corpus tools such as WordSmith and AntConc. WordSmith (Scott, 2008) is a proprietary concordancing tool for Windows (`http://www.lexically.net/wordsmith`). AntConc (Antony, 2005) is a freeware concordancing tool for Windows, Mac or Linux (`http://www.antlab.sci.waseda.ac.jp/software.html`). In both cases these tools are typically used on a PC with the corpus and the corpus tool locally installed. Their strong point is that users can easily collect their own corpora and process them with these tools.

A radically different approach enables corpus queries from ordinary web browsers. This has two major advantages: the user does not need to install special software, and the user does not need to store local copies of the corpora. A good example of a web-based interface to an annotated corpus is BNCweb (Hoffmann et al., 2008), a web interface for the British National Corpus. In BNCweb the front-end user interface runs in an orinary web browser and provides extensive facilities for querying the corpus, viewing concordances, and other services. The back-end MySQL database contains the British National Corpus, converted from its original XML format and indexed for fast processing with MySQL. However, BNCweb runs on conventional web servers, not in the cloud.

In earlier work (Wilcock, 2010) we described a prototype that demonstrated the use of language technology in a cloud computing environment. This version can be seen at `http://aelred-austen.appspot.com`. It runs on Google App Engine and presents a web browser inter-

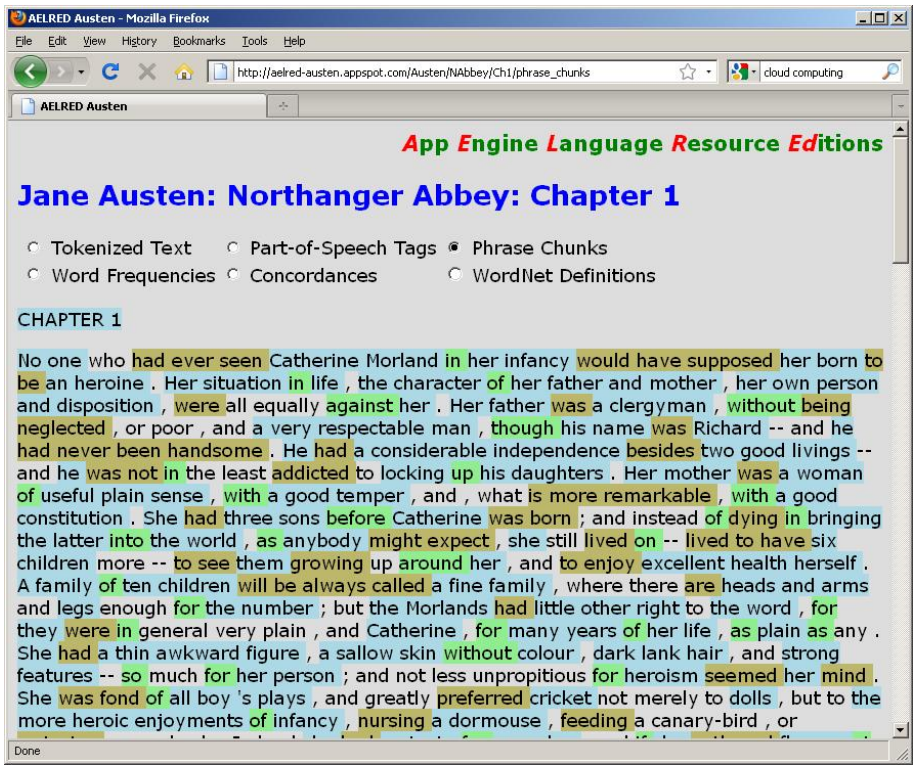

Figure 3: Example application: NP, PP, VP phrase chunks.

face to an annotated corpus of Jane Austen novels. The browser displays different types of annotations, including part-of-speech tagging, phrase chunks, and word sense definitions from WordNet. However, Wilcock (2010) did not address the problem of how to avoid lock-in to a proprietary framework. This is an important question that we discuss in Section 5.3..

4. An Example Application

Screenshots from the example application with the half-million-word annotated corpus of Jane Austen texts are shown in Figures 1 to 6.

Although we use NLTK tools for language processing as much as possible, the example application does not use the NLTK tokenizer `nltk.word_tokenize()` because there are specific problems in tokenizing the Gutenberg texts of the Jane Austen novels. One problem is the use of a double hyphen (`--`) to represent a dash. Wilcock (2010) gives an example from the third sentence in *Northanger Abbey* which includes the string `Richard--and`. This is tokenized as a single token by the standard NLTK tokenizer. Our example application therefore uses a regular expression tokenizer that splits this string correctly into three tokens. This can be seen in Figure 1.

The example application also does not use the NLTK part-of-speech tagger `nltk.pos_tag()` for the reasons given in Section 5.. The application uses an alternative pure Python tagger trained on the NLTK Treebank corpus, a subset of the full Penn Treebank corpus. The tagger is uploaded into App Engine as a `pickle` file. An example of text with part-of-speech tags can be seen in Figure 2.

Phrase chunks for NPs, PPs, and VPs are identified using NLTK's regular expression parser over POS tag sequences, and are annotated with IOB chunk labels. Phrase chunking is displayed with colour-coded highlighting as shown in Figure 3.

Simple word frequencies and concordances can also be displayed, as shown in Figure 4 and Figure 5. These are both rather basic, and certainly do not match the sophistication of dedicated concordance tools such as WordSmith, AntConc or BNCweb. The concordances are created using NLTK's `ConcordanceIndex()` method, and show all occurrences of a word in a novel, not chapter by chapter. The offsets for the whole novel are calculated off-line and uploaded to datastore in a serialized JSON format.

Words are also annotated with word sense definitions using NLTK's Python-WordNet interface. Words that have WordNet definitions are highlighted, and the definition pops up in a tooltip when the mouse hovers over the word, as shown in Figure 6.

The range of possible definitions for each word is restricted by the part-of-speech tag already decided by the POS tagger. A simple form of word sense disambiguation is used to select one definition to be displayed. This is based on the simplified Lesk algorithm, with the most frequent WordNet sense as back-off.

5. Technical Issues

This section discusses some potential problems relevant to our strategy and describes solutions. First, there are restrictions imposed by Google App Engine in order to support scalability. Next there are some technical issues in using

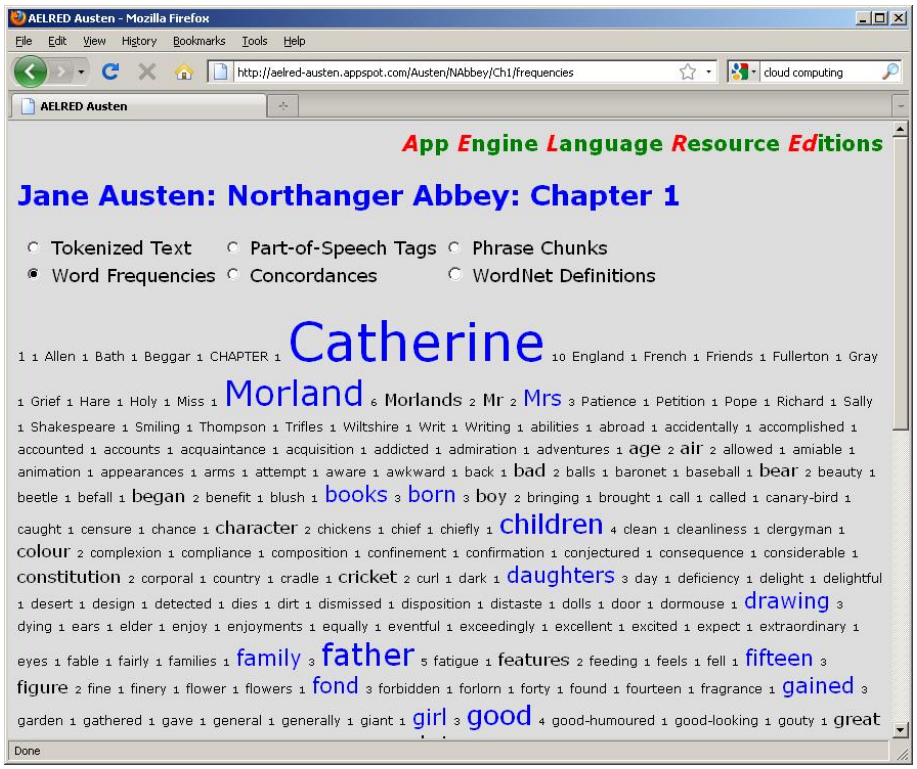

Figure 4: Example application: simple word frequencies.

specific NLTK tools with App Engine. Finally, there is the danger of lock-in to Google's proprietary framework.

5.1. Scalability and restrictions

When Google App Engine was designed, one of the key requirements was that it must allow massive scalability. As a result, small applications must be designed for scalability in the same way as large applications. To ensure scalability, various restrictions are imposed on all App Engine applications. There are different types of restrictions, on the programming language, maximum number of files, maximum file size, and so on.

A major programming language restriction is that the code must be pure Python, not depending on modules implemented in other language such as C. This means that you cannot upload code that uses numpy, which is written in C. You cannot use cPickle, but you can use pure Python pickle.

Up to now, the maximum file count in an App Engine application has been 3,000. If you bundle large packages (such as Django or NLTK) with your app, you could hit this limit. However, this problem can be avoided by using zipimport (Sanderson, 2008). In fact, recent versions of Django are included in recent versions of App Engine, so you do not need to bundle Django with your app, as (Sanderson, 2008) points out.

Up to now, the maximum file size allowed in App Engine has been 10 megabytes. In the NLTK version of WordNet, the file containing all the nouns is just over 15 megabytes, so the WordNet data cannot itself be uploaded into App Engine. Files can be annotated with WordNet definitions off-line, and the annotated files can be uploaded so long as they are less than 10 megabytes.

For the Jane Austen novels each chapter text fits easily within the maximum, and when annotations are added for part-of-speech tags and other small features, the file size is still less than the limit. However, when WordNet definitions are added the file size increases drastically because the definition strings are quite long and many words have multiple definitions, so some chapters can exceed the limit. This problem is solved by doing word sense disambiguation, so that only one definition is used.

5.2. NLTK and App Engine

NLTK includes a wide range of components implemented by different people in different ways, and some of them use numpy or other C modules. This means that you cannot simple do "import NLTK" in App Engine.

As (Wilcock, 2010) points out, there are two ways to use NLTK with App Engine. One way is to use NLTK off-line to create the required annotations. If the annotations are saved for example as JSON text files, these files can be included in the folders uploaded to the cloud as part of your App Engine app. This approach has the advantage that you can use all the NLTK components with no restrictions, even if they use C or numpy.

The other way is to make a stripped-down version of NLTK in a new folder, only including specific components that use pure Python. Then you can include this new NLTK folder in your app, and you can do "import NLTK".

In this approach, annotations are created by tools running inside the App Engine framework. As noted above, tools written in pure Python can be used in App Engine, but tools written in C cannot be used. Some of the NLTK tools are pure Python so they can be imported into App Engine suc-

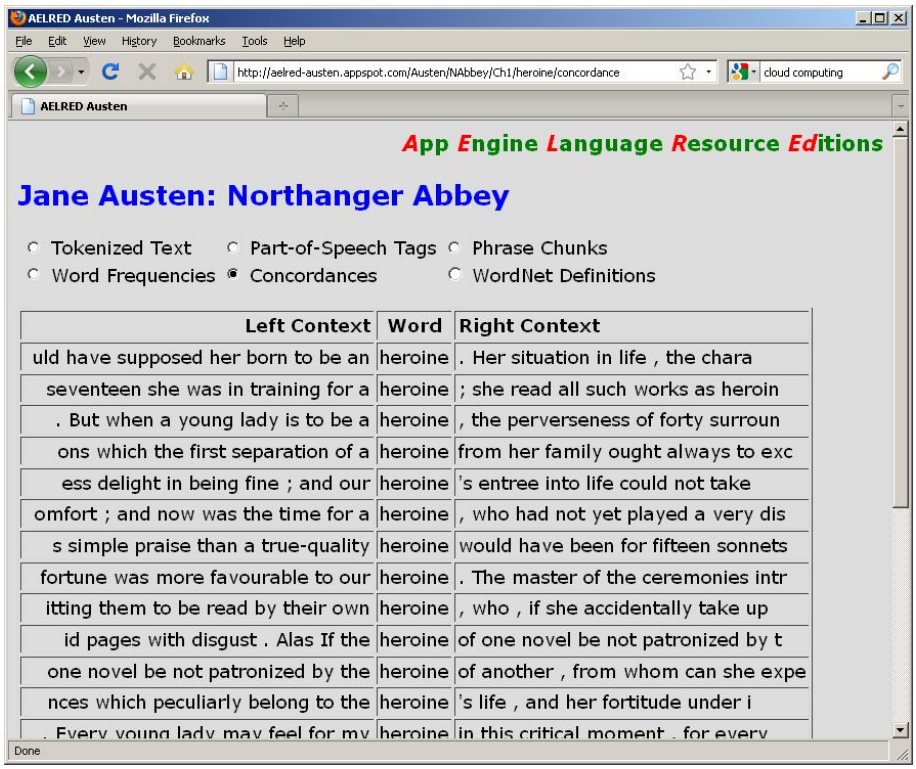

Figure 5: Example application: a simple word concordance.

cessfully, but some cannot. Alternative pure Python tools should be used.

Further details of which NLTK tools can and cannot be used in App Engine are discussed in (Wilcock, 2010).

5.3. Avoiding lock-in

There has recently been controversy about changes in the pricing scheme for commercial applications in App Engine, but free quotas are still available and in some cases the quotas have even been increased. While it is very attractive to run natural language processing applications and linguistic corpora free of charge on Google's infrastructure, there is always the possibility that charging might be introduced in the future. It is therefore advisable to beware of the danger of lock-in to one proprietary system, and even to have an exit strategy in case of need.

The danger of lock-in to Google's framework can be largely avoided by taking two steps. The first step concerns the web app front-end. By using a well-designed and widely-used open source web framework like Django, it will be much easier to move the application away from Google infrastructure to a more traditional server if that is desired in future, because standard servers can run standard Django web apps.

The second step concerns the back-end datastore. Although Django is normally used with standard SQL databases, Django's ORM (object-relational mapping) maps Python objects (logical models) to relations (database tables). This allows an SQL database to be used from Python code without actually writing SQL statements.

The open source `django-nonrel` project (Kornewald and Wanschik, 2011) is an extension of standard Django that maps Python objects at a higher level of abstraction, allowing either SQL databases or NoSQL databases to be used with the same models, provided the data models have not been designed around specific SQL-only or specific NoSQL-only features. This makes Django web apps portable between SQL databases and the App Engine datastore, thereby avoiding the danger of lock-in (Wanschik et al., 2010).

However, `django-nonrel` is not included in standard Django and is not supported by the Django Software Foundation. If using `django-nonrel` is considered unsuitable, there are two main alternatives.

One option is to keep the Django front-end and re-write the database back-end to use App Engine datastore explicitly. We did this for our example application and the conversion was very easy, as the ORM mappings for Django and App Engine are very similar. This version runs at `http://django-appeng.appspot.com`.

The other alternative is to use Django with a MySQL database and not with App Engine datastore. This avoids re-writing any code, and there are possibilities for running the application in the cloud. One option is to use the Google Cloud SQL service, which combines App Engine with a MySQL database. The pricing for Google Cloud SQL is not yet known, but the preview service is free. A version of our example application with Django and MySQL runs on Google Cloud SQL at `http://django-mysql.appspot.com`.

6. Conclusions and Future Work

The working example application shows that free storage and free delivery of annotated corpora can be achieved by

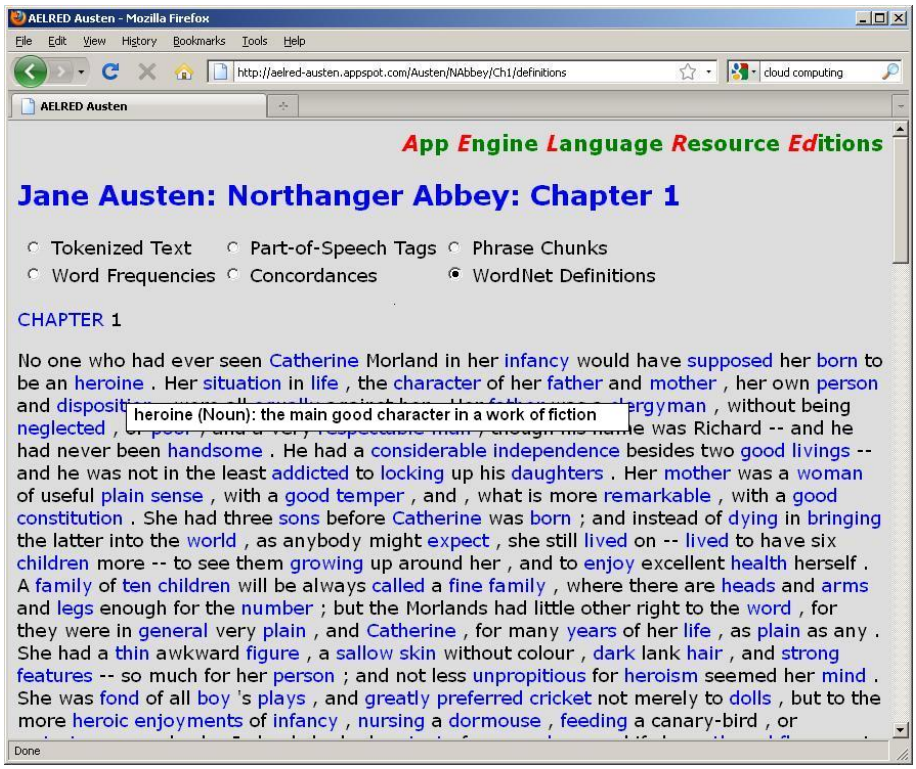

Figure 6: Example application: a word sense definition in a tooltip.

the approach described. Care must be taken to avoid lock-in to one proprietary infrastructure, but this risk can be minimized by adopting open source web frameworks like Django as basic components of the application.

Section 5.3. discussed approaches to avoiding lock-in. One option is to use Django with MySQL and not App Engine datastore, because MySQL can be used in a wide range of environments, either on cloud services or on conventional web servers. Using MySQL with the Google Cloud SQL service is currently free, but charging is expected later. There are other Platform-as-a-Service providers, such as Red Hat Cloud, offering free cloud services including Django and MySQL. We are currently setting up another MySQL-based instance of our example corpus application on Red Hat Cloud at `http://django-corpora.rhcloud.com`.

As we use JSON format rather than XML for the annotations (as mentioned in Section 2.5.), we are currently investigating document-oriented databases that use JSON format directly. These include CouchDB and MongoDB (which uses binary JSON: BSON). We are setting up a MongoDB-based instance of our example corpus application on Red Hat Cloud at `http://mongo-corpora.rhcloud.com`.

Future work will develop better methods for handling word frequency analysis and more sophisticated concordance queries, at least including multi-word phrases and part-of-speech tags. Several further corpora will be made available in the cloud, starting with the Brown Corpus which is nicely divided into small files ready for uploading and offers scope for genre-based concordance querying.

For this workshop, the most interesting future work would be to combine cloud delivery with crowd sourcing. App Engine has facilities for individual user authentication and for maintaining user-specific records in the datastore. If stand-off markup is used, updated annotations input by individual users could be stored as alternatives without damage to the exisiting annotations. Crowd sourcing algorithms could be deployed to decide which alternatives should be applied as updates to the displayed corpora. These possibilities await further work.

7. References

Laurence Antony. 2005. AntCon: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Proceedings of International Professional Communication Conference*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

Sebastian Hoffmann, Stefan Evert, Nicholas Smith, David Leed, and Ylva Berglund Prytz. 2008. *Corpus Linguistics with BNCweb - a Practical Guide*. Peter Lang, Frankfurt am Main.

Adrian Holovaty and Jacob Kaplan-Moss. 2009. *The Definitive Guide to Django (second edition)*. Apress.

Waldemar Kornewald and Thomas Wanschik. 2011. Django-nonrel - NoSQL support for Django. `http://www.allbuttonspressed.com/projects/django-nonrel`.

Dan Sanderson. 2008. Using Django 1.0 on App Engine with ZipImport. `http://code.google.com/appengine/articles/django10_zipimport.html`.

Mike Scott. 2008. WordSmith Tools version 5. Liverpool: Lexical Analysis Software.

Thomas Wanschik, Waldemar Kornewald, and Wesley Chun. 2010. Running Pure Django Projects on Google App Engine. `http://code.google.com/appengine/articles/django-nonrel.html`.

Graham Wilcock. 2010. Cloud computing for the humanities: Two approaches for language technology. In *Human Language Technologies - The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, Riga.

Guidance through the standards jungle for linguistic resources

Maik Stührenberg, Antonina Werthmann, Andreas Witt

Institut für Deutsche Sprache
R 5, 6-13
D-68161 Mannheim
{stuehrenberg|werthmann|witt}@ids-mannheim.de

Abstract

Research today is often performed in collaborated projects composed of project partners with different backgrounds and from different institutions and countries. Standards can be a crucial tool to help harmonizing these differences and to create sustainable resources. However, choosing a standard depends on having enough information to evaluate and compare different annotation and metadata formats. In this paper we present ongoing work on an interactive, collaborative website that collects information on standards in the field of linguistics as a means to guide interested researchers.

1. The importance of standards for collaborated resource development

Research today is often performed by teams of project partners from different institutions and countries. The first steps in such projects often focus on architectural issues, such as the choice of annotation formats or metadata standards. Project partners can only choose the best standards for their projects, however, when they have enough information to evaluate and compare standards. In this paper we will present ongoing work on an interactive, collaborative website that collects information on standards in the field of linguistics.

2. Different views on standards

Over the last 20 years, the annotation of linguistic phenomena has gone through a number of transitions, on both a general "meta" level and a more specific application-oriented level. First, meta languages such as SGML and later XML were established as standards. These two meta languages replaced the proprietary and binary formats that were used in annotation projects for linguistic data and were developed by the ISO/IEC (in case of SGML) and the W3C (in case of XML). Both organizations act in a wide field of specifications that may affect linguistic research, such as the W3C Recommendations XPath, XSLT, XML Schema or the Internationalization Tag Set (Lieske and Sasaki, 2007), or the ISO standards RELAX NG or Schematron. In addition, other general standards that are also crucial for language resources were developed by other organizations such as Unicode (The Unicode Consortium, 1991). These various specifications laid the groundwork for the application-oriented level, where initial steps were undertaken to harmonize the various efforts of linguistic researchers by developing a unified tagset for linguistic annotation. This was necessary since use of the same underlying meta language did not guarantee easy exchange of data or a sustainable use of the meta language (Stührenberg, 2008). One result of this movement was the Text Encoding Initiative (TEI) and its Guidelines. Development of TEI began in 1987 as an SGML application and the latest XML-based version, P5, was released in 2007 (Burnard and Bauman, 2007) and updated 2011 (Burnard and Bauman, 2011). It comprises 22 modules of over 520 elements and over 430 attributes, and allow for the annotation of various linguistic phenomena. Since the TEI is quite complex but has certain shortcomings regarding some linguistic theories, a third major transition regarding annotation of linguistic corpora is taking place.

There are already numerous specifications that deal with various aspects of linguistic annotation. Amongst these are the SGML-based Corpus Encoding Standard CES (Ide and Priest-Dorman, 1996; Ide, 1998), which has been developed within the Expert Advisory Group on Language Engineering Standards (EAGLES) as an application of the TEI P3 (Expert Advisory Group on Language Engineering Standards, 1996), and its XML-based successor XCES (Ide et al., 2000). Following the work of the EAGLES initiative, the ISLE (International Standards for Language Engineering) project, which has been carried out in collaboration between American and European groups under the Human Language Technology (HLT) programme within the EU-US International Research Co-operation, continued to develop and promote language technology standards, guidelines and tools (Calzolari et al., 2002).

Other annotation formats and frameworks have been developed through the course of several research projects, including the Potsdam exchange format for linguistic annotation (Potsdamer Austauschformat für Linguistische Annotationen, PAULA) (Dipper, 2005) or the Sekimo Generic Format (SGF) (Stührenberg and Goecke, 2008) and its successor XStandoff (Stührenberg and Jettka, 2009).

Since 2005, at least half a dozen efforts to standardize (technically, to create ISO standards for) various aspects of linguistic researches have been attempted. Among these specifications are the general Feature Structures (ISO/TC 37/SC 4, 2006) and the Linguistic Annotation Framework (ISO/TC 37/SC 4, 2011), the more specific Morpho-Syntactic Annotation Framework (ISO/TC 37/SC 4, 2008), the Syntactic Annotation Framework (ISO/TC 37/SC 4, 2010), and the Data Category Registry (DCR) (ISO/TC 37/SC 3, 2004), to name just the most prominent. The most recent (i.e. final) versions of these standards are usually not open and freely available on the Internet (although libraries often grant access to the public). Some information can be derived from scientific articles but these may already be out of date. Although most

of the standards mentioned above do relate to each other, the standardization process has no mechanism to coordinate standards, which may result in specifications becoming out of sync. Another practical issue is choosing which conceptual layer is covered by the standard (e.g. syntax, semantic, etc.).

2.1. Technical aspects

Technical questions such as the grammar formalism used or the notation can have direct consequences for choosing tools to process annotated resources. Some specifications deal with a single layer (such as the Morpho-Syntactic Annotation Framework and the Syntactic Annotation Framework), while others provide a general framework such as the Linguistic Annotation Framework. Others are not used for direct annotation at all; one example is the Data Category Registry, which should only be used as a registry for annotation standards concepts.

Most of the current annotation standards use the concept of standoff annotation introduced (Thompson and McKelvie, 1997) and discussed in the TEI as well. As a result, it is necessary to find/create annotation tools capable of dealing with the separation of content and markup, limiting the choice of tools that can be used to annotate resources – although one may observe that support for standoff annotation has increased in recent years (e.g. the web-based Serengeti annotation tool (Stührenberg et al., 2007), the Glozz Annotation Platform (Widlöcher and Mathet, 2009; Mathet and Widlöcher, 2011) or the newly developed Slate (Kaplan et al., 2011)).

2.2. Formal aspects

Among the formal aspects are the formal model, the constraint language used to define the markup language (and its respective expressive power), and the annotation model (inline vs. standoff). Although the formal model of an XML instance is that of a single-rooted tree, it is possible to encode graphs in XML as well (one has to differentiate between the XML instance as such which forms a tree and the language that is represented by it, which has no further restrictions). This can be achieved by using either quite general frameworks, such as the Linguistic Annotation Framework or Feature Structures, or by using meta markup languages, such as XStandoff.

The aspect of the constraint language used may be of interest regarding the expressive power of the markup language. This expressivity can be compared both in terms of technical features (such as data typing) and formal power. Both aspects have been subject to different research, e.g. (Murata et al., 2005) built up a taxonomy of schema languages which was refined by (Stührenberg and Wurm, 2010).

3. Providing Guidance

A large number of standards can be used in the creation of sustainable linguistic resources. Within the CLARIN-D project, the IDS is responsible for providing insight into various aspects of linguistic standards. The work presented in this paper aims to help interested researchers understand the relationship between various specifications and to choose the right standard for a given task. To support this, we are developing a lightweight and transparent taxonomy that can be used as an online guide for the most recent (and most prominent) specifications for language resources, especially annotation of linguistic data. This online guide will feature information addressing the issues raised here to help researchers differentiate between standards and choose the right one. It consists of two parts. The first part contains lightweight XML metadata descriptions of the various standards. This data is in the form of stripped-down markup language that can be easily modified with a text editor. The metadata is coded based on the TEI header, while the description of the features (including the aforementioned technical and formal aspects) is coded based on TEI's feature structures which in turn was standardized as (ISO/TC 37/SC 4, 2006). Following the distinction between technical and formal aspects we make assumptions about the meta language used (SGML vs XML), the constraint language that defines the markup language and the respective grammar class, and the notation (inline vs standoff), amongst others.

The second part takes these lightweight XML metadata descriptions as a knowledge base and allows the filtering of this data according to the different criteria stated above. The results can be transformed into different output formats that are readable by web browsers, and include textual and graphical representations.

The main parts of this system are the XML descriptions of annotation formats (or other standards), a database that stores the annotation instance (e.g. a native XML database) and a web frontend for both input and output using stylesheet transformation. The web frontend is designed to make the system useful for projects with many partners. We have currently completed both the XML annotation format and prototypic instances of the specifications' description (an excerpt is shown in Listing 1).

Listing 1: Example of a specification description

```
<spec xmlns:xsi="http://www.w3.org/2001/XMLSchema-
     instance" xml:id="SpecXces" topicRef="TopicGenAnn"
 xsi:noNamespaceSchemaLocation="http://localhost:8080/
      exist/apps/clarin/xsd/spec.xsd">
 <titleStmt>
  <title>XCES: Corpus Encoding Standard in XML</title>
 </titleStmt>
 <scope>Corpus annotation</scope>
 <description>
  <p>XCES is the XML version of the CES (Corpus Encoding Standard) ... </p>
  <!--[...]-->
 </description>
 <version xml:id="SpecXces104">
  <versionNumber>1.0.4</versionNumber>
  <date>2008-06-20</date>
  <respStmt>
   <resp>Editor</resp>
   <name type="person">Nancy Ide</name>
   <name type="person">Patrice Bonhomme</name>
  </respStmt>
  <features>
   <fs>
    <f name="metaLanguage">
     <symbol value="XML"/>
    </f>
    <f name="constraintLanguage">
     <symbol value="XSD"/>
    </f>
    <f name="grammarClass">
     <symbol value="LTG"/>
    </f>
    <f name="formalModel">
     <symbol value="Graph"/>
    </f>
    <f name="notation">
     <symbol value="Standt off"/>
```

```
    </f>
    <f name="multipleHierarchies">
     <fs>
      <f name="support">
       <binary value="yes"/>
      </f>
      <f name="item">
       <vColl>
        <string>standoff annotation</string>
       </vColl>
      </f>
     </fs>
    </f>
   </fs>
  </features>
  <address type="URL">http://www.xces.org/</address>
  <relation target="SpecCes" type="isVersionOf">
   <p>XCES is the XML instantiation of CES.</p>
  </relation>
 </version>
</spec>
```

The description of a specification can be subdivided into respective versions to distinguish different feature sets. For example, the P3 version of the TEI used the SGML meta language while form P4 onwards XML was used. However, while P4 used XML DTDs as constraint language the current P5 is based on RELAX NG. Since we only provide a small subset of any feature that can be relevant for a project, the description of the feature set is done via a TEI feature structure-like representation. Relations between specifications are described via the `relation` element. It contains two required attributes, `target` and `type`. While the former specifies the standard this one is related to, the value of the latter classifies the type of relation. We provide a list of relation types based on the DCMI Metadata Terms (DCMI Usage Board, 2010), such as *isApplicationOf* or *isVersionOf*, amongst others. DCR categories which can be obtained via ISOcat[1] could be used as well.

An even more lightweight format is used to store and describe the topics which are subsumed in a single XML instance.

Listing 2: Example of a topic description

```
<topics xmlns:xsi="http://www.w3.org/2001/XMLSchema-
     instance"
 xsi:noNamespaceSchemaLocation="http://localhost:8080/
      exist/apps/clarin/xsd/spec.xsd">
 <!--[...]-->
 <topic xml:id="TopicMetadata">
  <titleStmt>
   <title>Metadata</title>
  </titleStmt>
  <description>
   <p>Metadata contains information about other data ... </p>
   <!--[...]--A
  </description>
 </topic>
```

The format as such is defined by an XML schema description (XSD) because of XSD's strong data typing support. At present, the implementation shown above is stored into the native XML database eXist.[2] XQuery scripts transform the given information into different XHTML output files based on interactive web forms created with XForms (Boyer, 2009). Figure 1 shows a partial screenshot of the current implementation.[3]

A future incarnation will support a graphical overview of the relations between different specifications based on Scalable Vector Graphics (SVG), the one shown in Figure 1 contains a preliminary mockup. At the time of writing, the proposed system is a work in progress. The complete site will be launched, to coincide with the conference.

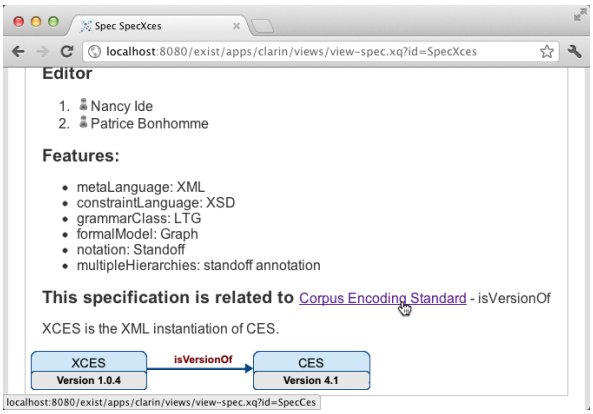

Figure 1: Partial screenshot of the current prototype.

4. Related approaches

There are already similar and related initiatives that try to help researchers deal with the variety of standards and language resources and that should be mentioned, although they cover a wider range of tools than our approach does. The LRE Map of Language Resources and Tools by FLaReNet (Fostering Language Resources Network)[4] and ELRA (European Language Resources Association) which was introduced at the LREC 2010 conference and collects information on both existing and newly-created language resources[5]. As a next step, the Language Library (Calzolari et al., 2011a) has launched for LREC 2012.

The FLaReNet Databook[6] comprises a picture of the current state of language resource technology and includes a practical orientation for the current standards landscape (Calzolari et al., 2011c; Monachini et al., 2011). Since the Databook states that information about standards has to be "constantly/periodically revised and updated by the community itself", we think that a open, web-based approach may be a means to this goal.

5. Outlook and further possible enhancements

Up until now, the relations between the specifications described are quite basic (cf. Section 3.). Possible future enhancements should not only address a more detailed graphical rendering of the relations but should enhance the type of relations as well, including mutually dependent relations between standards.

[1] See `http://www.isocat.org` for further details.

[2] See `http://www.exist-db.org` for further details.

[3] The prototype can be observed at `http://clarin.ids-mannheim.de/standards`.

[4] FLaReNet is a project initiative funded by the European Commission in the framework of the eContentplus Programme. See `http://www.flarenet.eu` for further details.

[5] A beta version can be found at `http://www.resourcebook.eu/LreMap/faces/views/resourceMap.xhtml`.

[6] Cf. `http://www.flarenet.eu/?q=FLaReNet_Databook` for further information.

6. References

John M. Boyer. 2009. XForms 1.1. W3C Recommendation, World Wide Web Consortium (W3C).

Lou Burnard and Syd Bauman, editors. 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. published for the TEI Consortium by Humanities Computing Unit, University of Oxford, Oxford, Providence, Charlottesville, Bergen, 10. Version 1.0.0. Last updated on October 28st 2007.

Lou Burnard and Syd Bauman, editors. 2011. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Published for the TEI Consortium by Humanities Computing Unit, University of Oxford, Oxford, Providence, Charlottesville, Bergen, 3. Version 1.9.1. Last updated on March 5th 2011.

Nicoletta Calzolari, Alessandro Lenci, Francesca Bertagna, and Antonio Zampolli. 2002. Broadening the scope of the EAGLES/ISLE lexical standardization initiative. In *COLING-02: Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization*, pages 1–8.

Nicoletta Calzolari, Riccardo Del Gratta, Francesca Frontini, and Irene Russo. 2011a. The language library: Many layers, more knowledge. In Calzolari et al. (Calzolari et al., 2011b), pages 93–97.

Nicoletta Calzolari, Toru Ishida, Stelios Piperidis, and Virach Sornlertlamvanich, editors. 2011b. *Proceedings of the IJCNLP 2011 Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, Chiang Mai, Thailand, 11. Asian Federation of Natural Language Processing.

Nicoletta Calzolari, Monica Monachini, and Valeria Quochi. 2011c. Interoperability framework: The FLaReNet action plan proposal. In Calzolari et al. (Calzolari et al., 2011b), pages 41–49.

DCMI Usage Board. 2010. DCMI Metadata Terms. DCMI Recommendation, Dublin Core Metadata Initiative, 10.

Stefanie Dipper. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin.

Expert Advisory Group on Language Engineering Standards. 1996. EAGLES Guidelines.

Nancy M. Ide and Greg Priest-Dorman. 1996. Corpus Encoding Standard (CES). Technical report, Expert Advisory Group on Language Engineering Standards (EAGLES).

Nancy M. Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Language Resources and Evaluation (LREC 2000)*, pages 825–830, Athens, 5. European Language Resources Association (ELRA).

Nancy M. Ide. 1998. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In *Proceedings of the First International Language Resources and Evaluation (LREC 1998)*, pages 463–470, Granada, Spain. European Language Resources Association (ELRA).

ISO/TC 37/SC 3. 2004. Terminology and other language and content resources — Data categories — Part 1: Specification of data categories and management of a data category registry for language resources. Committee Draft ISO/CD 12620-1 (N 509), International Organization for Standardization, Geneva, 7.

ISO/TC 37/SC 4. 2006. Language Resource Management — Feature Structures – Part 1: Feature Structure R. International Standard ISO 24610-1:2006, International Organization for Standardization, Geneva.

ISO/TC 37/SC 4. 2008. Language Resource Management — Morpho-syntactic annotation framework. Draft International Standard ISO/DIS 24611, International Organization for Standardization, Geneva.

ISO/TC 37/SC 4. 2010. Language Resource Management — Syntactic annotation framework (SynAF). International Standard ISO 24615, International Organization for Standardization, Geneva.

ISO/TC 37/SC 4. 2011. Language Resource Management — Linguistic annotation framework (LAF). Final Draft International Standard ISO/FDIS 24612, International Organization for Standardization, Geneva, 8.

Dain Kaplan, Ryu Iida, Kikuko Nishina, and Takenobu Tokunaga. 2011. Slate — a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, 26(2):89–101.

Christian Lieske and Felix Sasaki. 2007. Internationalization Tag Set (ITS). W3C Recommendation, World Wide Web Consortium (W3C), 4.

Yann Mathet and Antoine Widlöcher. 2011. Stratégie d'exploration de corpus multi-annotés avec glozzql. In Mathieu Lafourcade and Violaine Prince, editors, *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, pages 143–148, Montpellier. Association pour le Traitement Automatique des langues (ATALA).

Monica Monachini, Valeria Quochi, Nicoletta Calzolari, Núria Bel, Gerhard Budin, Tommaso Caselli, Khalid Choukri, Gil Francopoulo, Erhard Hinrichs, Steven Krauwer, Lothar Lemnitzer, Joseph Mariani, Jan Odijk, Stelios Piperidis, Adam Przepiorkowski, Laurent Romary, Helmut Schmidt, Hans Uszkoreit, and Peter Wittenburg. 2011. The Standards' Landscape Towards an Interoperability Framework: The FLaReNet proposal Building on the CLARIN Standardisation Action Plan, 7.

Makoto Murata, Dongwon Lee, Murali Mani, and Kohsuke Kawaguchi. 2005. Taxonomy of XML Schema Languages Using Formal Language Theory. *ACM Transactions on Internet Technology*, 5(4):660–704.

Maik Stührenberg and Daniela Goecke. 2008. SGF – an integrated model for multiple annotations and its application in a linguistic domain. In *Proceedings of Balisage: The Markup Conference*, volume 1 of *Balisage Series on Markup Technologies*, Montréal, Québec, 8.

Maik Stührenberg and Daniel Jettka. 2009. A toolkit for multi-dimensional markup: The development of SGF to XStandoff. In *Proceedings of Balisage: The Markup Conference*, volume 3 of *Balisage Series on Markup Technologies*, Montréal, Québec, 8.

Maik Stührenberg and Christian Wurm. 2010. Refining the

Taxonomy of XML Schema Languages. A new Approach for Categorizing XML Schema Languages in Terms of Processing Complexity. In *Proceedings of Balisage: The Markup Conference*, volume 5 of *Balisage Series on Markup Technologies*, Montréal, Québec, 8.

Maik Stührenberg, Daniela Goecke, Nils Diewald, Irene Cramer, and Alexander Mehler. 2007. Web-based annotation of anaphoric relations and lexical chains. In Branimir Boguraev, Nancy M. Ide, Adam Meyers, Shigeko Nariyama, Manfred Stede, Janyce Wiebe, and Graham Wilcock, editors, *Proceedings of the Linguistic Annotation Workshop*, pages 140–147, Prague.

Maik Stührenberg. 2008. Sustainability of Text-Technological Resources. In Andreas Witt, Georg Rehm, Thomas Schmidt, Khalid Choukri, and Lou Burnard, editors, *Proceedings of the LREC 2008 Workshop "Sustainability of Language Resources and Tools for Natural Language Processing"*, pages 33–40. ELRA/ELDA.

The Unicode Consortium. 1991. The Unicode Standard. Version 1.0, Volume 1. Technical report, The Unicode Consortium, Reading, MA.

Henry S. Thompson and David McKelvie. 1997. Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe '97: The next decade – Pushing the Envelope*, pages 227–229, Barcelona.

Antoine Widlöcher and Yann Mathet. 2009. La plate-forme glozz : environnement d'annotation et d'exploration de corpus. In Mathieu Lafourcade and Violaine Prince, editors, *Actes de la 16e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009) – Session posters*, Senlis. Association pour le Traitement Automatique des langues (ATALA).

Supporting Collaborative Improvement of Resources in the Khresmoi Health Information System

Lorraine Goeuriot[1], Allan Hanbury[2], Gareth J. F. Jones[1], Liadh Kelly[1], Sascha Kriewel[3], Ivan Martinez Rodriguez[4], Henning Müller[5], Miguel A. Tinte[4]

[1]Dublin City University – Dublin, Ireland
lgoeuriot, gjones, lkelly@computing.dcu.ie
[2]Technische Universitaet Wien –Vienna, Austria
allan.hanbury@tuwien.ac.at
[3]Universitaet Duisburg-Essen – Duisburg, Germany
sascha.kriewel@uni-due.de
[4]Atos Origin Sociedad Anonima Espanola – Madrid, Spain
ivan.martinez, miguel.tinte@atos.net
[5]University of Applied Sciences Western Switzerland – Sierre, Switzerland
henning.mueller@hevs.ch

Abstract

Since medical knowledge relies on both scientific knowledge and real-life experience, the importance of user contributions to improve resources in health systems cannot be underestimated. We present work from the Khresmoi project which aims to develop a multilingual multimodal search and access system for biomedical information and documents. Khresmoi targets three distinct user classes with differing levels of medical knowledge and information requirements, namely: general public, general practitioners, and, as an example of an area of clinical expertise, radiologists. The Khresmoi system will provide these users with valuable (whose quality has been evaluated and approved) and enriched (meta information from biomedical knowledge bases is added) medical information, selected to fit their medical knowledge and their preferred language. The system will include novel collaborative components of the system are designed to provide means for users to contribute to the system's knowledge by adding or correcting annotations to the documents, as well as a collaborative platform where they will be able to share their own files and both annotate and discuss them.

1. Introduction

Annotation of biomedical data is vital in order to be able to organise and structure the knowledge it contains, and to select and deliver information relevant to the information need of a searcher seeking to address a medical information need from these sources. In this paper, we describe our current work exploring how users (e.g. patients, physicians, etc.) of a medical system can help to improve it by contributing to the quality of its resources and by adding their knowledge to the stored information.

This work is being carried out within the Khresmoi project[1], which aims to develop a multilingual and multimodal search and access system for biomedical information and documents (Hanbury et al., 2011). The Khresmoi project is being targeted at three groups of end users: two groups with general medical interests (general public and general practitioners) and a group of clinicians with specialised expertise (radiologists); all speaking different languages, having different medical knowledge levels and differing levels of knowledge of the languages of the target documents. The system is based on a library of valuable medical documents (images and text) that are enriched using a medical ontology such as UMLS[2] (Unified Medical Language System) or MeSH[3] (Medical Subject Headings) and knowledge bases such as the LinkedLifeData[4]. The Khresmoi system is being designed to enable our users to correct computed knowledge (meta information and translations), as well as share their experience.

Based on a collection of biomedical documents, including medical 2D images and 3D volumes, automatically annotated with biomedical ontologies, we plan to provide users with the potential to correct errors in these automatic annotations. Since medical knowledge relies both on scientific knowledge and experience, medical literature may not be enough to understand a treatment, a procedure or even the description of a disease. Document meta-information and comments from users can help gathering that knowledge in a single space. For example, a young radiologist will have to check different resources and maybe colleagues to spot an area of interest on an X-ray image. With such a system, he will be able to search for similar images and then use the meta-information/annotations to validate his diagnosis. we will also provide them with tools to share their knowledge through notes and comments on documents. Both the user and the system can benefit from such collaborative tools: improving the quality of data will improve quality of the medical system search, and sharing knowledge and experience helps physicians in their everyday practice. The system will also provide automatic translations of the queries and documents. As automatic translation methods do not give perfect results, we will allow users to correct transla-

[1] http://khresmoi.eu/

[2] http://www.nlm.nih.gov/research/umls/

[3] http://www.nlm.nih.gov/mesh/MBrowser.html

[4] http://linkedlifedata.com/

tion errors as well.
The next section describes related work in medical related collaboration tools. Section 3. provides an overview of the Khresmoi project and its objectives, along with a description of the project's user interface system and resources used. Section 4. describes how users can collaborate to improve the system resources by updating annotations and translations, as well as communicate through comments and discussion threads. Finally Section 5. summarises the paper and outlines the focus of our ongoing work.

2. Related Work

Collaboration by editing digital resources to correct and augment their content is key to obtaining richer information. Knowledge, especially in such specialised domains as medicine, relies on scientific knowledge and experience. However, gathering knowledge from text sources by using automated information extraction methods only produces partially correct scientific knowledge of the data due to errors in the extraction process, and will generally be much less reliable than human-annotation. Web 2.0 technologies enable users to collaborate in the development of content, and an inclination do to this has been observed in the medical domain (Eysenbach, 2008). *Ask Dr Wiki*[5] and *Medpedia*[6] are two well-known wikis where physicians can create content, and collaborate on its editing. These wikis must provide complex validation systems in order to guarantee the quality of the information published. The purpose of these websites is mainly to improve online health information. Another online collaborative annotation tool, called *Brat*, provides a user-friendly interface to display and change annotations on text from a web browser. Registered users can view and annotate online files and upload their own files. It has been used for BioNLP extraction tasks and is mainly natural language processing (NLP) focused (Stenetorp et al., 2012). Collaborative projects have also been defined for particular communities of practice, where users sharing patients or interests can discuss cases, information and even manage meetings. For example, the SOMWeb system (Falkman et al., 2008) assists the community of Swedish oral medicine practitioners. Using OWL (Web Ontology Language) to model their data, it allows users to add cases, notes, discussions and manage community aspects.
Medical wikis provide users with a way to gather their knowledge in creating new content, while community of practice collaborative systems are specific software or online systems allowing collaboration in a very specific framework. However, none of these systems provides access to other resources, which is one of the main uses of the Internet. The time practitioners can spend online is rather limited: they spend on average less than 5 minutes to answer a question (Hoogendam et al., 2008). Expecting them to be active on different platforms is unrealistic. A system providing all these services at the same time would be valuable. The Khresmoi system, presented in this paper, is designed to provide a search service on valuable and enriched medical documents. The system includes collaborative components intended to enable users to improve resource documents and engage in discussions.

[5] http://askdrwiki.com/
[6] http://www.medpedia.com/

3. Khresmoi System

The Khresmoi project aims to develop a multilingual multimodal search and access system for biomedical information and documents. Khresmoi is adopting a user-centred approach to designing medical information search tools, for which three groups of end users are defined. Two of these are groups with general medical interests: general practitioners and members of the general public. The Khresmoi system is intended to provide them with innovative text search features to interrogate the huge amount of medical information available, including that appearing in journals, websites, Wikipedia and clinical guidelines. These users wish to rapidly find answers to their queries that are suitable for their level of expertise. The other user group that Khresmoi focuses on is radiologists, as an example of clinicians with a specific expertise. For radiologists we plan to provide advanced image search to support them in their work. The Khresmoi system is being developed within a four year project which is now in the first half of year two. During the first year of the project, the requirements of the end users were obtained through surveys and interviews. Following this, the design process for the Khresmoi system has led to a specification of: the characteristics of the target user groups, the types of search tasks that the users would perform, the resources that each user type wishes to access, and the search tools and refinements needed by each user type to carry out their tasks. An interesting result of the survey is the perceived importance of the collaborative aspects of search for medical professionals, who wish to see their peer's opinion on documents and also additional examinations that can increase their confidence in a diagnosis.

3.1. Khresmoi Users and Their Needs

In this section we summarise the surveys carried out within the Khresmoi project to investigate what the different categories of users need from a health information system.

- 385 members of the general public, mostly highly educated and coming from healthcare (not physicians) and IT backgrounds answered the survey. They came from 42 European countries (with the highest numbers of contributors coming from France and Spain). The most researched topics by these users are: general health, chronic diseases and lifestyle. When they were asked what were the most important characteristics of search tools, they mentioned the relevance and trustworthiness of the results.

- 556 physicians and 4 final-year medical students, mostly Internet savvy and with regular patient contact were surveyed. These respondents came mainly from Austria, Switzerland and the United Kingdom. The topics they search on the most are: drugs, treatments and medical education. Currently they mostly use generic search engines (such as Google). Specialist physicians also search for clinical trials and have a preference for medical research databases or society

websites, whereas general practitioners also search for disease description and tend to use more general health websites.

- 34 radiologists were surveyed, a majority of them young subjects currently with little radiology experience, however several of them had more than 15 years experience. They came mainly from Switzerland and Austria. Image search (search for images matching certain disease or body parts) was mentioned as a common task, but time consuming (often more than 10 minutes) and with 65% success in completing the task. One of the main points of a search tool is then to be able to find good and relevant image results quickly. Subjects would also like to be able to upload an image on a search tool as a query, to find similar images of similar cases.

From these surveys, we can see that the quality of the information, as well as its relevance and trustworthiness are very important criteria for every kind of user. Medical practitioners and radiologists mentioned the need to share information: medical practitioners wanted to have access to a secured community where they can exchange information about cases and share or update their knowledge; and radiologists mentioned that feedback from colleagues on past/current cases was valuable information. Therefore, users express the need of high quality information, as well as interactivity and communication functionalities. More details on these surveys can be found in the public deliverables of the Khresmoi project: (Pletneva and Vargas, 2011) for general public, (Gschwandtner et al., 2011) for medical professionals and (Müller, 2011) for radiologists.

Web2.0 and social media are having an impact on the medical domain, both on the specialist side (Giustini, 2006; Eysenbach, 2008) and on the patient side (Fox, 2011). This change has raised concerns about the quality of information (Denecke and Nejdl, 2009): without any editorial process, how can it be guaranteed? However, Web2.0 is subject to a "socially Darwinian process" (also called *positive network effect*): (Boulos et al., 2006) note with regard to wikis that "because of [the] openness and rapidity that wikipages can be edited, the pages undergo an evolutionary selection process not unlike that which nature subjects to living organisms".

If the user contributions are done in a controlled and secured way, with an adapted moderation system, the quality of information can still be guaranteed. What we propose here within Khresmoi is to let the users directly contribute to the quality of the information by correcting the metadata (annotation and translation of multilingual content), as well as to be able to freely comment and discuss cases in a secure environment.

3.2. Khresmoi Resources

The Khresmoi system will potentially index a very large number of documents from the biomedical domain. As the collection is a very long process, we gathered datasets for our first prototype in order to observe specific users behaviour. To improve the search, the approach of annotating the documents with entities important in the medical domain is being adopted, where the entities are taken from a knowledge base of domain ontologies in the medical and life sciences, such as the LinkedLife Data (semantic data integrationplatform for the biomedical domain).

Datasets used within the project for the first year prototype include 2D and 3D images, as well as text. The 3D image collection consists of: realistic clinical data (medical images and reports from the Vienna University Hospital, constituting over 3 TeraBytes of data) and lung data (medical images and reports collected in the University Hospital of Geneva, corresponding to more than 100 interstitial lung disease cases). These two collections have been anonymized and annotated using RadLex[7] and MeSH.

The 2D image collection is a collection from ImageCLEF2011 (Kalpathy-Cramer et al., 2011). It contains 231,000 images from the PubMed Central Database and corresponding articles, with articles annotated with MeSH. The text collection gathers MEDLINE[8] abstracts, UMLS[9] definitions, a set of Health on the Net[10] classified documents about diabetes. All these documents have been annotated with LinkedLife Data. These datasets have been designed for the first Khresmoi prototype and will be extended as part of the ongoing work of the project.

For the text annotation work during the project, extensive use of manual feedback from professional annotators is made to correct the annotations, and hence allow the system to improve the automated annotation through learning. However, the extensive use of professional annotators is not a sustainable approach, and the system will have to increasingly rely on annotation corrections from the end users. For the cross-lingual search, use of resources for which translated versions of terms are linked to each other is made, such as the MeSH thesaurus[11].

4. Collaborative Plans in Khresmoi

In this section, we describe technologies that have been developed within the project and our development plans for the future of the project.

During the first year of the Khresmoi project, a user interface framework based on ezDL technology has been developed. We are currently extending this to implement our plans to create tools to enable users to collaborate. An evaluation phase of these components is planned later in year 2 following their development.

4.1. EzDL System

The user interface of the Khresmoi system is based on ezDL[12], the successor of the Daffodil software (Fuhr et al., 2002) developed at the University of Duisburg-Essen. EzDL is a multi-agent search system for heterogeneous data sources and a tool-set for building search user interfaces to support complex tasks. It allows for simultaneous searches in multiple digital libraries through a unified interface and

[7] http://www.radlex.org/
[8] http://www.ncbi.nlm.nih.gov/pubmed/
[9] http://www.nlm.nih.gov/research/umls/
[10] http://www.hon.ch/
[11] http://www.ncbi.nlm.nih.gov/mesh
[12] http://www.ezdl.de/

query syntax, and presents a merged and enriched view of the results. The tools provided by ezDL allow users to work with the results and can be arranged in customizable perspectives.
EzDL is composed of a server part consisting of a directory and a large number of agents, and clients that contain a selection of loosely-coupled tools which serve as a user interface to the system (see Figure 1).
The server-side agents connect to the search and query support services provided within Khresmoi, handle user authorisation, user profile management, logging, storage of user data and queries, and the caching of documents. Two basic clients are available within Khresmoi: a search desktop written in Java (see Figure 2), as well as a browser application that uses Java Server Faces. Users can either search as guests or obtain a personal account. A personal account allows for a persistent search history spanning multiple search sessions and offers access to a document depository called 'personal library', where a user can store found and uploaded documents, as well as favourite queries and authors, and categorise them with personal tags.
An account will also be necessary to contribute to the system's knowledge by adding or correcting annotations on the documents. Guests and registered users alike can use the search tool with query formulation support which offers spelling corrections and disambiguation of medical terms. The results are presented in a combined list that searchers can group using options like date, type of document (e.g. image or text), category of document (e.g. treatments, symptoms, genetics) or audience (e.g. general public, practitioners or researchers). The search tool also offers filtering, sorting by different criteria, and export options. Documents that have already been inspected, stored, printed or otherwise handled by the user are clearly marked with icons in the result list. The detail tool of ezDL offers a preview of documents from the result list or from the personal library. It shows document metadata (authors, publication date, publication type, journal or conference), annotations of the content and summaries where available. A link to the full document (website, article or media file) is also provided.

4.2. Khresmoi Collaborative Components Development Plan

As mentioned in Section 3.1., surveyed potential users expressed the desire to share knowledge, especially medical professionals and communities of practice. Web2.0 facilitates this knowledge sharing on the web by allowing users to directly contribute information (e.g. Wikipedia or Medpedia). The Khresmoi system will provide users two ways to share their knowledge:

- correction of existing annotations and translations created by the system;
- creation of comments on Khresmoi documents or on documents uploaded by users, that can target a specific part of the document (region of interest in an image or sentence/paragraph in a text) or the whole document.

These two collaborative approaches will improve Khresmoi resources by adding: explicit knowledge through corrections, and implicit knowledge through comments. While the system can directly benefit from explicit knowledge, both can be useful for users. As mentioned in the surveys (see Section 3.1.), the quality and the relevance of a search result are very important criteria. If users can correct resources on the system that they are also using to get information, they can directly benefit from their input: better translations and annotations improve the quality and relevance of the documents (e.g. though ranking process). We also observed in the user surveys that experience sharing played an important role in physicians and radiologists everyday practice. This system could allow them to do it online, with colleagues that can be in other institutions. For example, a radiologist could give feedback through notes on a radiological image to a general practitioner who needs advices. Physicians can share comments on new clinical trials with other physicians or highlight useful recommendations in a document for patients.
We provide details on these collaborative approaches in Sections 4.2.1. and 4.2.2..

4.2.1. Users Correcting Annotations and Translations

To improve resources in the Khresmoi system users will be able to update and correct errors in such resources while using the system. This can take several forms: direct correction of errors or omissions in annotation or translation, or manual contribution of new knowledge, e.g. translations, or verification or clarification of automatically extracted suggested updates for resources. In addition to supporting users in updating resources in operation, we will also explore methods such as collaborative editing to keep the resources up to date. From a technical perspective we propose the development of a *Collaborative Resources Framework* to support the improvement process. Figure 3 presents an overview of the *Collaborative Resources Framework* as well as the external communication with other components of the Khresmoi system.
We can distinguish two types of processes in the context of collaborative improvement: updating and validating. These processes are aligned with components in the collaborative framework: the *Resource Updater* is responsible for the annotation and translation management; and the *Validator* responsible for managing the life cycle of the user annotations and translations. Both annotations and translations will be by default in a *Pending* state and could change to a *Validated* or a *Refused* state. We next describe these two processes in greater detail.

Resource Updater : This component will manage annotation and translation updates incoming from the ezDL user interface. It consists of two main subcomponents: *Annotation Manager* and *Translation Manager*. The *Annotation Manager* is responsible for implementing the workflows for *New Creation* and *Update Annotation* functionalities as they are offered by ezDL. The Annotation Manager will insert annotations, and updated annotations, into the User Profile database. The Annotation Manager also writes to/reads from an *Annotation State Store*. This store manages the different possible states associated with annotations (*Pending,*

Validated and Refused). The default annotation status will be *Pending*, requiring a user to validate the annotation and change the status to *Validated* or *Refused*. The *Translation Manager* will implement the functionality associated with the *Update Translation* workflow in Figure 3. To fulfil this task, this component will use the *Multilingual Translation Framework* (MTF) provided by our system. The MTF controls the management and storing of translations and user translation updates, hence they will be stored outside of the Collaborative Resources Framework. The *Update Translation* functionality will be provided by the ezDL user interface and the manager will recover the translation from the MTF. Similar to annotations, translations will require user validation. The status associated with user translations will represent their validation status.

Validator : This component will provide the functionalities needed for managing the life cycle associated with annotations and translation. As mentioned previously, when a user adds or updates one concrete document annotation or translation the *Resource Updater* marks as *Pending* the state of the annotation or translation. To support this functionality the Updater will use the *Annotations State Store* for annotations and the MTF for translations. The Validator component will allow users to carry out two types of actions over pending annotations or translations: validate or refuse them. Following user validation, the Validator component commits or discards the annotation/translation as appropriate.

4.2.2. Users adding comments to documents

As we said previously, medical professionals' knowledge is based on scientific knowledge but also relies strongly on their experience. While the scientific knowledge can be more or less similar across persons and available in books and online, experience is rather individual. For this reason it is very important and interesting for practitioners to share this knowledge. Our system aims to provide users with a simple system to share their knowledge and experiences. Registered users will be allowed to share documents from the project library and add comments and discussions on these files. They will also be able to upload their own files (e.g. patients report or x-ray radiography) to the system, which will be anonymous (no patient information) and private (the user will choose people to share the file with). Users can add comments on the whole document or on a region of interest (Figures 4 and 5).
Users' rights fall into 3 categories:

Read : Users will be allowed to read comments from other users. The comments will be accessible for users within the same category (general public, medical practitioners or radiologists), unless the author specifies other categories (e.g. a physician could highlight an interesting paragraph for patients).

Write : Users will be allowed to create/write new comments. Whatever the document is, these users will be allowed to add new comments to discuss it or add new knowledge (annotations).

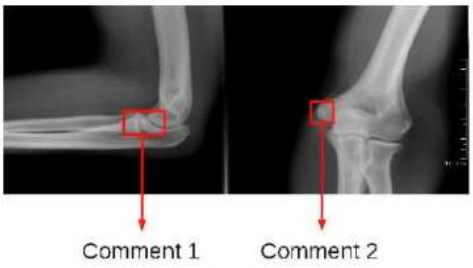

Figure 4: Example of annotations on an image (from `http://en.wikipedia.org/wiki/Radiography`)

Figure 5: Example of annotations on a text (from `http://en.wikipedia.org/wiki/Radiography`)

Modify : Users will be able to edit or delete all their comments. They will not be allowed to modify other users comments.

All users, even if they are not registered, will be allowed to read comments written for their category. Registered users will be able to write new comments and modify their own comments. When a new comment is added, the user will have to choose categories of users allowed to read it (e.g. a doctor can write comments for patients). Registered users will be able to edit or delete their own comments.

4.3. Evaluation of the Collaborative Components

Empirical and user-centered evaluation strategies have been developed for the Khresmoi system, which will be conducted in the coming months. The user-centered part of this system evaluation strategy encompasses evaluation of the collaborative components using target user groups. This will entail subjects from each category of user using the system to fulfil predefined scenarios. Feedback gained on the collaborative components through these evaluations will be used to adapt the components to make them more user-friendly and suitable to user practice.

5. Summary and Ongoing Work

In this paper, we have presented a set of collaborative functionalities that will be included in the Khresmoi medical information search system. This system will provide users with a valuable search tool for medical documents that are available in multiple languages, and enriched using medical thesauri. Medical documents can be processed by the system using information extraction tools to include semantic annotations. To do this, a knowledge base of domain ontologies in the medical and life sciences is used. The system will also provide automatic translation of the queries and documents, and provide users with facilities to collaborate to correct these annotations and translations. User collaboration will also be possible through a component which will allow users to add comments and start discussions on documents from the library or their own files. The development of these components is ongoing. These components, along with the system, will be evaluated in the coming months, through both empirical and user-centered evaluations. Patients, medical practitioners and radiologists will partake in the controlled user-centered system evaluations. The system will be improved based on feedback from these evaluations. Following this, the system will be deployed for use by real users. Among other things this will allow us to both assess the quality and value of users' input, and investigate how user input could further contribute to the system. For example, comments and discussions from physicians on a document describing a case might provide rich information that the system could learn to process.

6. Acknowledgements

This research described in this paper has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n. 257528 (KHRESMOI)

7. References

Maged Kamel Boulos, Inocencio Maramba, and Steve Wheeler. 2006. Wikis, blogs and podcasts: a new generation of web-based tools for virtual collaborative clinical practice and education. *BMC Medical Education*, 6:41.

Kerstin Denecke and Wolfgang Nejdl. 2009. How valuable is medical social media data? content analysis of the medical web. *Journal of Information Sciences*, 179:1870–1880.

Gunther Eysenbach. 2008. Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness. *Journal of Medical Internet Research*, 10(3).

Göran Falkman, Marie Gustafsson, Mats Jontell, and Olof Torgensson. 2008. Somweb: A semantic web-based system for supporting collaboration of distributed medical communities of practice. *Journal of Medical Internet Research*, 10(3).

Susannah Fox. 2011. Health topics. Technical report, Pew Research Center, February.

Norbert Fuhr, Claus-Peter Klas, André Schaefer, and Peter Mutschke. 2002. Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In *Research and Advanced Technology for Digital Libraries. 6th European Conference, ECDL 2002*, pages 597–612, Heidelberg. Springer.

Dean Giustini. 2006. How web 2.0 is changing medicine. *British Medical Journal*, 333(7582):1283–1284.

Manfred Gschwandtner, Marlene Kritz, and Celia Boyer. 2011. D8.1.2: Requirements of the health professional search. Technical report, Khresmoi Project, August.

Allan Hanbury, Célia Boyer, Manfred Gschwandtner, and Henning Müller. 2011. Khresmoi: towards a multilingual search and access system for biomedical information. In *Med-e-Tel*, Luxembourg.

Arjen Hoogendam, Anton F.H. Stalenhoef, Pieter F. de Vries Robbé, and John P.M. Overbeke. 2008. Answers to questions posed during daily patient care are more likely to be answered by uptodate than pubmed. *Journal of Medical Internet Research*, 10(4).

Jayashree Kalpathy-Cramer, Henning Müller, Steven Bedrick, Ivan Eggel, Alba García Seco de Herrera, and Theodora Tsikrika. 2011. Overview of the CLEF 2011 medical image classification and retrieval tasks. In *CLEF 2011 working notes*.

Henning Müller. 2011. D9.1: Report on image use behaviour and requirements. Technical report, Khresmoi Project, May.

Natalia Pletneva and Alejandro Vargas. 2011. D8.1.1. requirements for the general public health search. Technical report, Khresmoi Project, May.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012 (to appear)*.

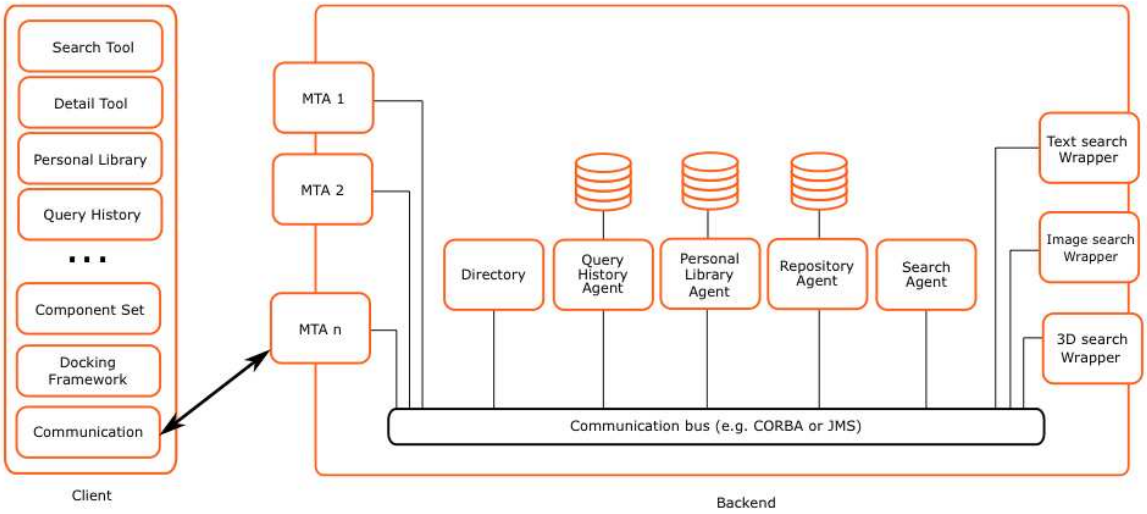

Figure 1: Architecture of ezDL

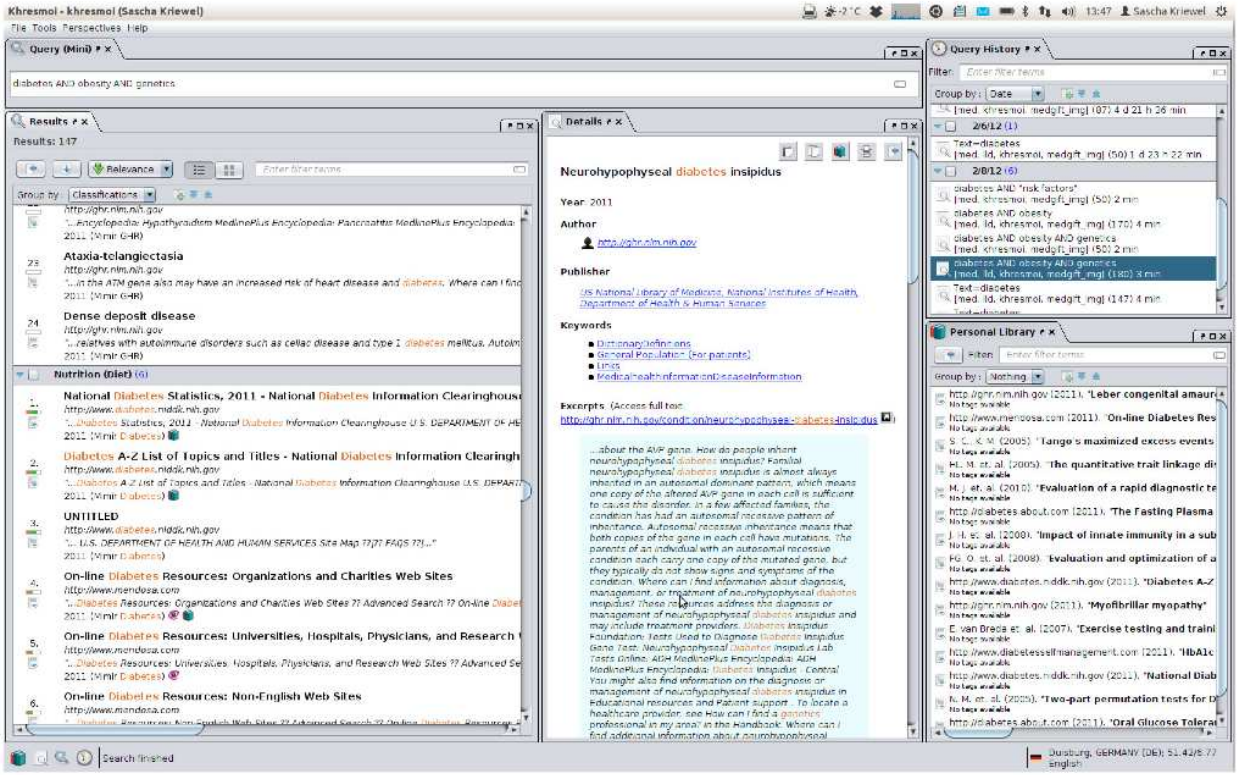

Figure 2: Screenshot of ezDL Interface

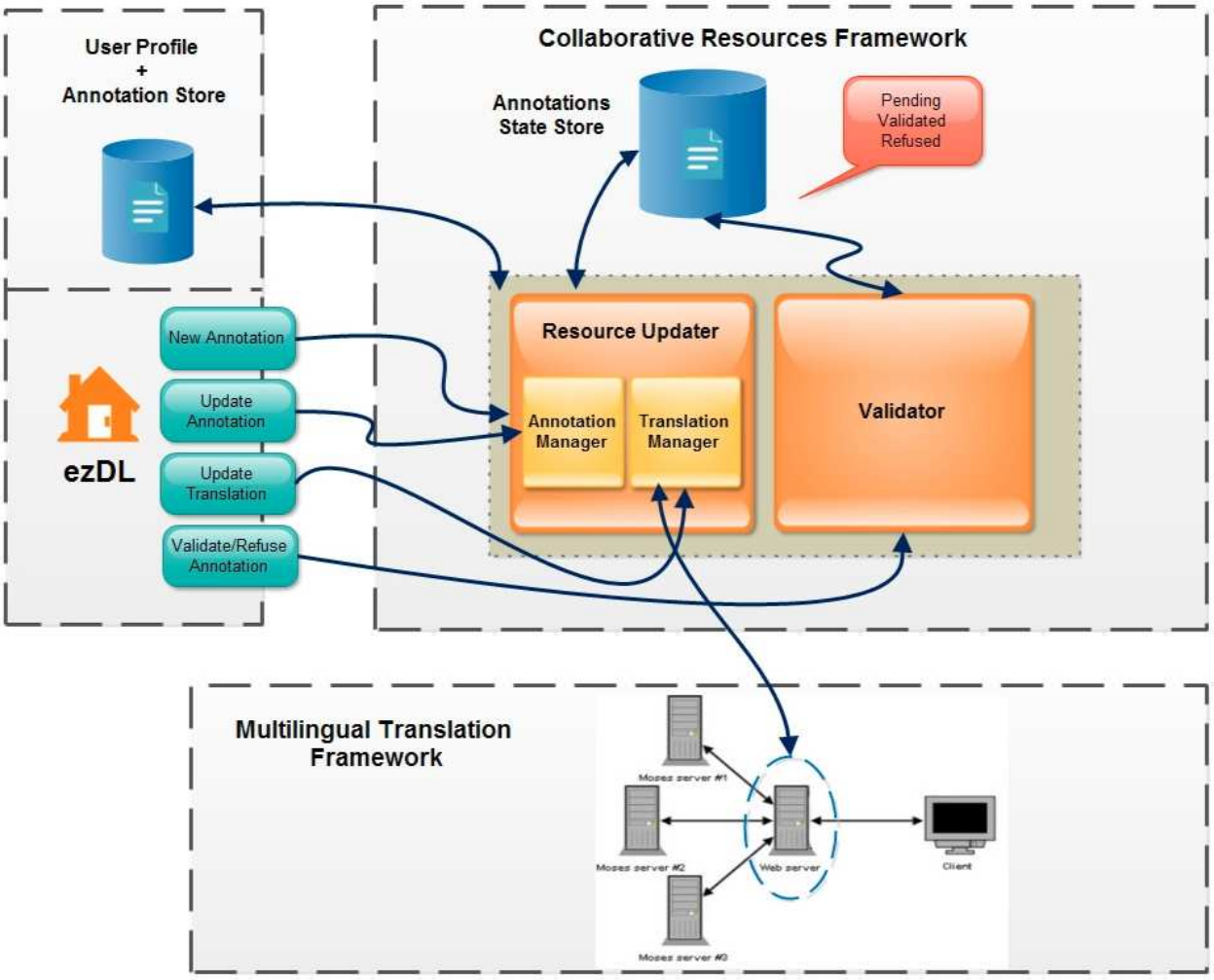

Figure 3: Collaborative Resources Framework Architecture

Building parallel corpora through social network gaming

Nathan David Green

Charles University in Prague
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
green@ufal.mff.cuni.cz

Abstract

Building training data is labor-intensive and presents a major obstacle to the advancement of Natural Language Processing (NLP) systems. A prime use of NLP technologies has been toward the construction machine translation systems. The most common form of machine translation systems are phrase based systems that require extensive training data. Building this training data is both expensive and error prone. Emerging technologies, such as social networks and serious games, offer a unique opportunity to change how we construct training data. These serious games, or games with a purpose, have been constructed for sentence segmentation, image labeling, and co-reference resolution. These games work on three levels: They provide entertainment to the players, the reinforce information the player might be learning, and they provide data to researchers. Most of these systems while well intended and well developed, have lacked participation.

We present, a set of linguistically based games that aim to construct parallel corpora for a multitude of languages and allow players to start learning and improving their own vocabulary in these languages. As of the first release of the games, GlobeOtter is available on Facebook as a social network game. The release of this game is meant to change the default position in the field, from creating games that only linguists play, to releasing linguistic games on a platform that has a natural user base and ability to grow.

Keywords: Serious Games, Social Networks, Parallel Corpora, Collaborative Games

1. Introduction

Parallel corpora are sentence aligned translations from multiple languages. These texts are often hand curated and professionally translated and as of recently, have been automatically harvested from the Internet. These may take the form of bilingual news releases (Nadeau and Foster, 2004), multilingual Wikipedia entries (Adafre and de Rijke, 2006), or boot strapping approaches that form comparable corpora, corpora that are not necessarily sentence align word for word translation (Munteanu and Marcu, 2005).

Phrase based machine translation along with many other machine learning methods make great use of parallel corpora. The outputs of these systems may be sentence translation, a sentence disambiguation, or simply a multi-lingual dictionary.

Crowd sourcing, such as mechanical turk (Kittur et al., 2008; Callison-Burch, 2009), and collaborative games such as image labeler (von Ahn and Dabbish, 2004), Duolingo (Savage, 2012), PackPlay (Green et al., 2010), Phrase Detectives (Chamberlain et al., 2009) and Lgame.cz (Hladká et al., 2009) have all shown the ability to reduce the cost and effort needed from single individuals to construct annotations for linguistic data. (Hladká et al., 2011) gives a more complete survey of this area of research. None of these systems to our knowledge have looked at the creation of parallel corpora in a game setting. This is a logical platform since many people learning multiple languages are often in school and of the typical game playing age and would look for software that might aid them in their learning.

There are many social gaming platforms on the market these days. From mobile devices such as Android, RIM and iOS, to social networking sites such as Google+ and Facebook. For this study we will focus on Facebook due to its market share and stable API for social networking and a common history of making social games. Continuing the theme of the paper, future releases should focus on the mobile phone market, to broaden the reach of the linguistic communities efforts in datamining.

2. Game setup

GlobeOtter is set up as a common application for multiple games. As a player gathers more points, new games become available to them. This allows us to keep the players' interest while also controlling which games are appropriate at different levels of expertise shown by the player. Figure 1 shows one of the intro screens in which the player can choose between the two current games which are unlocked. Along the side of the game are tabs that allow the players to access different areas of the software. These areas include functionality to let players add photos from their Facebook albums into the game along with submitting a translation in the language of their choosing. As with many games we also include a leader board and personal stats so they players know how they perform in comparison to their friends and other players. GlobeOtter gives players rank "levels" depending on their percentile in a particular language. So for instance, a player could say they are in 90th percentile in Indonesian but only the 50th percentile for English. The player can post this information to their wall to show off and to also recruit other players to add translations and captions in multiple languages to their photos. We feel that players will be highly motivated to see what captions and what languages people use to describe their photos. The two initial games, Translation Game and Word Fall Game are describe below.

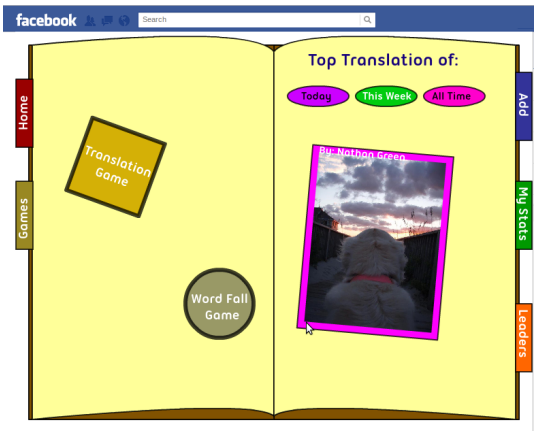

Figure 1: The first game screen in which the user can select which game to play and view the post popular translation which encourages them to submit their own.

2.1. Translation Game

Translation Game, as pictured in Figure 2, is a game intended to help players broaden their vocabulary in new languages. At the beginning of the game the player selects two languages, one they know and one they want to know. One of the languages they select they will read and one they will write. The player can choose whether to read or write the language they are learning. Given that the user knows one reference language, we eliminate one word for the "written" language. The player must then type the correct word in. Then the user will then receive a score for the answer and their answer is stored in our database. Since these are user submitted translations they may or may not be correct. For this reason we have had to come up with a new scoring technique, outlined in 2.1.2..

2.1.1. Questions

For each language pair we supply 10 question and photo pairs to get started. All questions and photos from that point on are submitted by players from their Facebook photo albums. We think this kind of personalization and game "fame" will keep Facebook users coming back for more game sessions. The accuracy of the player supplied questions and answers will be detected based on how other players do on those questions. If a set of players commonly translate a question differently then the supplied answer we have the ability to substitute that question answer pair with one of the more commonly agreed on answers.

2.1.2. Scoring

We allow players to submit the question answer pairs and generally there is no one correct answer to a translation. For this reason a new scoring system was needed since we can't evaluate an answer as correct or incorrect. The scoring is based out of 100 points. 25 of those points are awarded the player if the player matches the original translation. 75 of those points are based on what percentage of other players gave the same answer. This allows the player to get more points if the community agrees on a different answer than the original author gave.

Figure 2: Example of a user playing the Translation game.

2.2. Word Fall Game

Word Fall Game focuses on learning the word order in a given sentence. The game, using a physics engine, drops words from the sky and the user has to grab them and throw them in the proper direction so that they land in a basket. There is a basket for each word. If a word is placed in the wrong order, in this case the wrong basket, then it is shot back out into the sky for the user to throw again. When all words are placed in the correct order in each respective basket, the player receives a score for that sentence. Figure 3 shows a typical game screen.

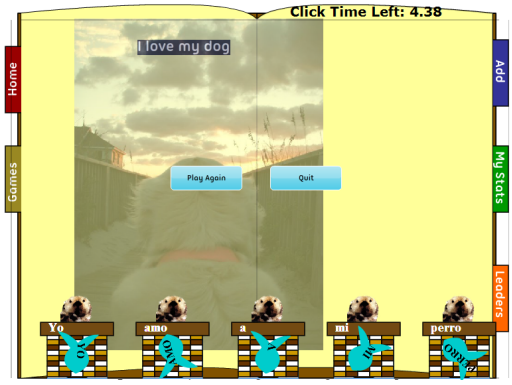

Figure 3: A typical game screen from Word Fall Game in which the player has to rearrange the word order as the words (fish) fall from the sky.

In contrast to the Translation Game, this game only focuses on word order. The player still sees a reference translation in a language they are familiar with and the reference photo displayed in the background. Since no typing is required in this game the player can focus solely on the reordering of words.

The questions and scoring are much simpler for this game. The questions are the same phrases and photos that were supplied by players in the Translation Game. This allows us an extra mechanism to verify that a translation is probable. The scoring is done by how long a player holds down a mouse button to reorder words. The less they use the mouse the more points they get. While the scoring does not directly impact the player's language learning, we feel the need to "throw" words around to get more points is more

entertaining for the player.

2.3. Data Generated

The data we are generating is completely user supplied and therefore does not have any legal issue with distribution. Data generated will be made available in publicly released sets at www.globeotter.com. Data for each language pair is stored in a database and is queryable. In later iterations, image and term pairs will be released to aid information retrieval and object detection tools. The parallel corpora is composed of the sentences supplied by the players about a particular photo. Since we will likely have multiple translations that are generally agreed on by fellow players, the data will have an additional benefit of multiple reference translations, which is rare with parallel corpora.

Due to the nature of social networking and photos available through these sites, we expect the genre of the text to be of a casual nature. The photos should most likely be made of a combination of travel photos and pictures of friends. The informal nature of social networks should allow for training data that will be more applicable to translating blogs, emails, and instant messages, instead of the typical government, financial, and news based training data that is typically available.

2.4. Research Questions

While the mere generation of new parallel sentences might be useful to the community at large, the new research questions that appear from such work may be equally beneficial. First, both games are intended to be used by players who may in all likelihood be monolingual and just starting to work on a second language. Because of this we expect many translations to be incorrect. While having a second game available to "check" the accuracy of the first will help, if most players are similar in expertise, they may just verify errors. To address this, we think the best solution is a hidden pre-test in which the player is tested on known translations. This cannot happen until some sentences are manually checked by experts. In such a situation a player's submissions can be weighted in a fashion similar to their bleu (Papineni et al., 2002) score against the gold data. Players can also be compared against each other using inter-annotator agreement metrics (Helmreich et al., 2004).

A second questions is how to train a statistical machine translation system effectively when we have a possibility of N reference translations where N is the number of players in the system. It is unlikely to approach N but realistically the number could be much higher than the 1-4 reference translations seen by systems today. Using this data might provide fruitful for judging systems trained on other data. Along with many criticisms, a common problem with automated scoring techniques is the lack of ability to recognize equally valid but sentences that are very different in terms of n-gram composition (Zhang et al., 2004). In this case it would be likely that a subset of the reference translations could verify the appropriateness of such sentences. Combining this with the player accuracy scores listed above could provide very useful information to the community.

3. Conclusion

We have presented a new game framework, GlobeOtter, that is linguistically motivated and allows for the automatic creation of parallel corpora from game data. At the time of writing this paper we have support for 6 languages, but we intend to add new virtual keyboards and languages as players wanting to learn those languages add the game on Facebook. We will release parallel training data in pre-compiled sets at regular intervals as more players join the game. It is our hope that changing the platform for which serious games are released will have an dramatic and immediate effort on the amount of open and free linguistic data available. Applying such a game to inexperienced translators raises many research questions in data filtering but also changes the traditional view of statistical machine translation which traditionally has one reference translation to possibly a system that can use a large number of references to better an overall translation.

4. Acknowledgements

This research has received funding from the European Commission's 7th Framework Program (FP7) under grant agreement n° 238405 (CLARA)

5. References

S. F. Adafre and M. de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore, August. Association for Computational Linguistics.

Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2009. Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 57–62, Suntec, Singapore, August. Association for Computational Linguistics.

Nathan Green, Paul Breimyer, Vinay Kumar, and Nagiza Samatova. 2010. Packplay: Mining semantic data in collaborative games. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 227–234, Uppsala, Sweden, July. Association for Computational Linguistics.

S. Helmreich, D. Farwell, B. Dorr, N. Habash, L. Levin, T. Mitamura, F. Reeder, K. Miller, E. Hovy, O. Rambow, and A.Siddharthan. 2004. Interlingual annotation of multilingual text corpora. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 55–62, Boston, MA, May 2 - May 7.

Barbora Hladká, Jiří Mírovský, and Pavel Schlesinger. 2009. Play the language: Play coreference. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*,

pages 209–212, Suntec, Singapore, August. Association for Computational Linguistics.

Barbora Hladká, Jiří Mírovský, and Jan Kohout. 2011. An attractive game with the document: (im)possible? *The Prague Bulletin of Mathematical Linguistics*, (96):5–26.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *The twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456, New York, NY.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31:477–504, December.

David Nadeau and George Foster. 2004. Real-time identification of parallel texts from bilingual newsfeed.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ. Association for Computational Linguistics.

Neil Savage. 2012. Gaining wisdom from crowds. *Commun. ACM*, 55(3):13–15, March.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 319–326, New York, NY.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system. In *In Proceedings of Proceedings of Language Resources and Evaluation (LREC-2004*, pages 2051–2054.

Three steps for creating high-quality ontology lexica

John McCrae, Philipp Cimiano

Cognitive Interaction Technology Center of Excellence, University of Bielefeld
Universitätsstraße, D-33615 Bielefeld, Germany
{jmccrae,cimiano}@cit-ec.uni-bielefeld.de

Abstract

Sophisticated NLP applications working on particular domains require rich information on both the linguistic properties of words and the semantics of these words. We propose a three-step methodology for the creation of high-quality ontology-lexica which combine detailed syntactic information with deep semantic information about words and their associated meanings. Our proposed methods consists of three steps: first we rely on a standard NLP pipeline to create a preliminary version of the ontology lexicon automatically. In this step, the automatically created lexicon is linked to existing legacy lexical resources. The second step involves referencing existing lexical and semantic resources and importing data. Finally, a manual review step is required that is supported by a novel editor to facilitate the inspection and manual validation and modification and thus continuous refinement and improvement of the ontology lexicon.

1. Introduction

For many sophisticated NLP applications, such as question answering (Unger and Cimiano, 2011), natural language generation and machine translation (Beale et al., 1995; McCrae et al., 2011a), which work with text in specific domains, the creation of domain-specific lexical resources. However, the process of creating such resources often involves a significant amount of manual effort. In this paper, we propose a three-step method for creating *ontology-lexica* (Cimiano et al., 2007; Peters et al., 2007). Ontology lexica essentially specify how words, phrases etc. should be interpreted in the context of a given domain ontology and are thus crucial for ontology-based NLP applications. In particular, we propose the creation of ontology lexica by firstly creating an initial resource using a fully automatic process that builds in part on statistical natural language processing techniques for aspects such as identification of part-of-speech. Secondly, the process refers to existing resources and includes extra information from these sources in a semi-automatic manner, consulting the user primarily when ambiguity exists. Finally, the process involves a manual review of the results, allowing the user to correct errors that may have been introduced by the automatic tools. In such a way we envision that a resource such as an ontology-lexicon can be created quickly and easily for specific domains. We present this work in reference to a system called *lemon source* that allows for the creation of ontology-lexica using the *lemon* (McCrae et al., 2012a) ontology-lexicon format.

2. Ontology-lexicon models

Ontologies are widely used to represent semantics and the OWL format (McGuinness et al., 2004) has provided a standard format that has lead to the creation of a large number of ontological resources on the web, creating the *Semantic Web*. These ontologies have been applied to a number of natural language processing tasks. However, as noted by Buitelaar et al. (2009), the linguistic information contained in ontologies is typically not sufficient for NLP applications. Thus, in the past we have proposed the *lemon* model for formalizing and representing lexica which enrich ontologies with information about how the ontology elements are realized in different natural languages. *lemon* builds on existing work on Semantic Web lexical resources, in particular the LexInfo (Cimiano et al., 2011) and the LIR models (Montiel-Ponsoda et al., 2008) as well as the lexicon modelling framework, LMF (Francopoulo et al., 2006). The model thus aims to provide a richer description of lexico-linguistic information related to classes, properties and individuals in the ontology. In particular the *lemon* model contains a core set of elements describing a path between the ontological entity and the (string) label ("core path") consisting of **lexical entries** uniquely identified by URIs and available on the web as RDF data (Lassila et al., 1998). Lexical entries themselves consist of **lexical forms**, which record the inflectional variants of an entry and **lexical senses**, consisting of a **reference** to the ontology. A lexical form may have multiple **representations** in different scripts and/or orthographies and may have different phonetic representations and a lexical sense may be further described by pragmatic constraints.

In addition, the *lemon* model has a number of modules that extend the core path to handle the linguistic data required by these applications, in particular the following modules are used:

- **Linguistic Description**: Allowing for elements to be assigned to linguistic categories, e.g., of gender, case, number.
- **Variation**: Allows elements within the ontology-lexicon to be linked to elements of the same or other ontology-lexica.
- **Phrase structure**: Description of the decomposition of terms into other terms.
- **Syntax and Argument Mapping**: Consisting of syntactic frames and their correspondence to semantic predicates in the ontology.
- **Morphology**: Compact representation of form variants for highly synthetic languages.

Further details are described in the *lemon* cookbook[1].

3. Methodology

We propose a methodology for the creation of ontology-lexica as a three-step process illustrate each of the steps by an example.

3.1. Automatic resource creation

The first step we apply is to use automatic tools to create a skeleton resource that we can further enrich at the later stage. The methodology for this was described in (McCrae et al., 2011b) and we will recap it here. As input we assume that we have a resource that contains a set of terms for the domain, such as the labels for an ontology in OWL[2]. As an example, currently our system applies the following sub-steps:

- **Tokenization of multi-word terms**: This step involves analysing the label to see if it consists of multiple words. For European languages this is achievable with simple finite state automata, but it is often more complex for languages such as Chinese, Japanese or Korean (e.g., Wu and Fung (1994)).
- **Part-of-speech detection**: The next step is then to apply a part of speech tagger to deduce the part-of-speech of the word(s) in the label. In particular we use the Stanford Tagger (Toutanova and Manning, 2000).
- **Stemming**: We then normalize inflected forms of words in our label by means of a stemmer. For English we use the Stanford Tagger's stemmer and for other languages the Snowball stemmer[3].
- **Decompounding**: For some languages, notably German and Dutch, we need to break up compound words into their individual words, for example breaking "Qualitätsverbesserungskommission" ("quality improvement committee") into "Qualität", "Verbesserung" and "Komission." This is in practice achieved by applying the stemmer multiple times using the Viterbi algorithm.
- **Parsing**: After this we apply a parser to deduce the structure of a phrase if it consists of multiple words. In particular we use the Stanford Parser (Klein and Manning, 2003).
- **Frame detection**: We also detect for a verb or relational expression (such as "capital of") the number and kind of arguments it can take. Currently this is performed using rules based on the phrase structure/part-of-speech of the term, but could also be achieved by corpus analysis.
- **Subterm detection**: We search for common subterms across multiple input terms, extract these subterms from them and introduce new lexical entries for these subterms.
- **Term variation**: Here we use syntactic rules to suggest variants for terms; for example, we find in English that terms of the form $\texttt{NN}_1$ $\texttt{NN}_2$ can often also be expressed as $\texttt{NN}_2$ `of the` $\texttt{NN}_1$ such as "prostate cancer" and "cancer of the prostate."
- **Semantic relation induction**: We can use the lexical form of a word to induce semantic relationships between the terms. Currently, we induce hypernym relationships if two terms are subterms of one another, for example "personal profile document" is a type of "document."

Our system currently supports English and German, with partial support[4] for French, Dutch, Spanish and Chinese, which we plan to increase to full support.

3.2. Semi-automatic resource re-use

After having created a preliminary version of the ontology lexicon using automatic processing, we proceed to link this resource to other external resources. In particular we use two kinds of resources: machine-readable dictionaries, which have already been aligned to the *lemon* model (McCrae et al., 2012c) and semantic resources we find from the Web. More specifically, we use two lexical resources: WordNet (Fellbaum, 2010) and Wiktionary[5]. We rely on the following criteria to search for possibly aligned terms in the resources:

- The canonical (lemma) form is the same
- The part-of-speech is the same if present
- The two entries do not have contradictory values for a property, e.g., different grammatical genders
- The entries do not have a contradictory inflected form, e.g., a different plural form

It was found in (McCrae et al., 2012c) that 21.6% of WordNet entries could be matched to Wiktionary pages using this method of which 97.2% were unambiguous (in that there was only a single candidate Wiktionary page); the remaining WordNet entries has no equivalent in Wiktionary. This shows that the overlap between WordNet and Wiktionary is rather low in general.

The second kind of resource we attempt to link to are semantic resources, which we discover by using semantic web source engines such as the Watson search engine (d'Aquin et al., 2007). We detect similar concepts in these resources using a vector space model alignment algorithm similar to the one described by Trillo et al. (2007).

For both methods, we automatically link the ontology-lexicon to the relevant resource if it can be done so unambiguously. However, if multiple candidates are found by

[1] `http://lexinfo.net/lemon-cookbook.pdf`

[2] We note that there is significant effort required to identify the terminology required for a specific task, however automatic methods for extracting a term from a domain can be used.

[3] `http://snowball.tartarus.org`

[4] Generally, a trained model for the parser or tagger is not available

[5] `http://www.wiktionary.org/`

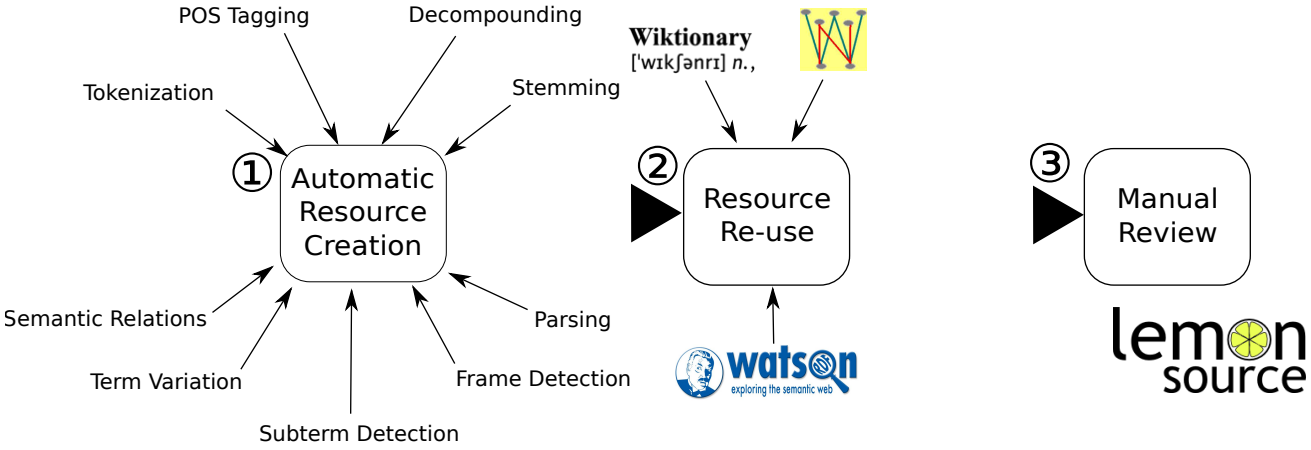

Figure 1: The three step procedure for creating an ontology-lexicon

the linking procedure we consult the user to allow them to select which resource should be used.

3.3. Manual review

The final step in the creation of an ontology-lexica is to manually review the result. While this could be achieved by examining the serialised (i.e., XML) form of the resulting resource, this would not be practical as many users, especially those with a non-technical background, would have trouble understanding the form of the resource, and it may be difficult to keep track of the progress. For this reason we have created a web application we call *lemon source* (McCrae et al., 2012b) that displays the result of the automatic and semi-automatic extraction and allows the user to edit the resulting entries in an intuitive manner. In addition, the system also allows the creation of meta-data about the ontology-lexicon, in particular assigning each entry to a number of statuses, such as "for review", "accepted" or "rejected." Lemon source allows users to collaboratively work on the lexicon by allowing their shared, remote use and by making updates made by one client immediately available to all clients, such as in Cunningham et al. (2003).

4. Discussion

This methodology for the creation of ontology-lexica is necessary for the sophisticated ontology-based NLP applications that we target, as we find that neither automatic methods, existing resources nor manual resource creation are sufficient to meet the challenge of creating high-quality lexica.

In the case of automatic ontology-lexicon induction, we would ideally hope that the result would be accurate and sufficient for the nature of the tasks we envision. However, while the systems we use have very high accuracy, they cannot said to be perfect. Our system achieves between 99.1% and 81.5% precision depending on the ontology (see (McCrae et al., 2011b)). A fundamental issue with the creation of a language resource by automatic methods is that any text processing system that uses an automatically generated language resource could achieve at least as good performance by directly integrating the tools used to create the language resource. As an example of this, consider that our text processing system needs to know the part-of-speech of terms used within a phrase; a language resource could extract this using a part-of-speech tagger, as we do. However, a statistical part-of-speech tagger will produce more information, such as the probability of individual words being tagged with a particular part-of-speech and other potential candidate taggings, which could be utilized by the end system. It is of course possible that we could include such information in the language resource at the risk of needlessly bloating the language resource in a manner that could make it difficult to use in practise. Nevertheless, materializing this information into an ontology-lexicon has the potential to reduce costs overall as people interested in exploiting an ontology for a given NLP application could download and reuse existing lexica instead of creating them from scratch.

In the case of reusing existing language resources, we have a clear advantage in that we can assume that these resources are of very high quality and much less likely to contain errors. However, there is a clear issue that for domain terminology it is highly unlikely that the resources contain all required entries. This proves to be significant when applying text processing applications that are dependent on language resources to new domain. However, much of the necessary information for these applications cannot easily be deduced by automatic methods, especially the extraction of specific relations between concepts and relationships involving multiple concepts (Zhou, 2007).

Finally, manual editing systems are ultimately necessary for the creation of high-quality resources. However, the creation of a complex language resource is extremely time-consuming and often requires users with specific training in linguistic resources. Moreover, it has been shown that complex annotation schemes like those required for structured resources like ontology-lexica lead to a lot of errors (Butler et al., 2000). As such, reducing the complexity of the scheme and the amount of the resource that needs to be created is a key goal of the manual annotation (Bayerl et al., 2003), and this can be carried out by incorporating automatic assistance (Smith et al., 2008).

5. Conclusion

We have proposed a three-step methodology for the creation of high-quality ontology lexica based on the use

of automatic tools, semi-automatic re-use of existing language resources and manual review, and presented a detailed overview of *lemon source*, an implemented web application that supports this methodology.. Each of these steps is extremely valuable for creating such resources, but the single steps have significant and complementary costs. Thus, by combining all three methodologies, high quality language resources which have high coverage and high accuracy for particular domains, can be quickly created.

6. References

P.S. Bayerl, H. Lüngen, U. Gut, and K.I. Paul. 2003. Methodology for reliable schema development and evaluation of manual annotations. In *Proceedings of the Workshop on Knowledge Markup and Semantic Annotation at the Second International Conference on Knowledge Capture (K-CAP 2003)*.

S. Beale, S. Nirenburg, and K. Mahesh. 1995. Semantic analysis in the Mikrokosmos machine translation project. In *Proceedings of the 2nd Symposium on Natural Language Processing*, pages 297–307.

P. Buitelaar, P. Cimiano, P. Haase, and M. Sintek. 2009. Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, pages 111–125.

T. Butler, S. Fisher, G. Coulombe, P. Clements, I. Grundy, S. Brown, J. Wood, and R. Cameron. 2000. Can a team tag consistently?: Experiences on the Orlando project. *Markup Languages*, 2(2):111–125.

P. Cimiano, P. Haase, M. Herold, M. Mantel, and P. Buitelaar. 2007. LexOnto: A model for ontology lexicons for ontology-based NLP. In *Proceedings of the OntoLex07 Workshop at the 6th International Semantic Web Conference*.

P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. 2011. LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

H. Cunningham, V. Tablan, K. Bontcheva, and M. Dimitrov. 2003. Language Engineering Tools for Collaborative Corpus Annotation. In *Proceedings of Corpus Linguistics 2003*, pages 80–87.

M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. 2007. Watson: a gateway for the semantic web. In *Proceedings of the 4th Annual European Semantic Web Conference*.

C. Fellbaum. 2010. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243.

G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. 2006. Lexical markup framework (LMF). In *Proceedings of the 2006 International Conference on Language Resource and Evaluation (LREC)*, pages 233–236.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

O. Lassila, R.R. Swick, et al. 1998. Resource description framework (RDF) model and syntax specification. Technical report, W3C Recommendation.

J. McCrae, M. Espinoza, Montiel-Ponsoda, G. Aguado-de Cea, and P. Cimiano. 2011a. Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In *Fifth workshop on Syntax, Structure and Semantics in Statistical Translation (SSST-5)*.

J. McCrae, D. Spohr, and P. Cimiano. 2011b. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC)*.

J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. 2012a. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*.

J. McCrae, E. Montiel-Ponsoda, and P. Cimiano. 2012b. Collaborative semantic editing of linked data lexica. In *Proceedings of the 2012 International Conference on Language Resource and Evaluation (LREC)*.

J. McCrae, E. Montiel-Ponsoda, and P. Cimiano. 2012c. Integrating wordnet and wiktionary with lemon. In *Workshop on Linked Data in Linguistics 2012*.

D.L. McGuinness, F. Van Harmelen, et al. 2004. OWL web ontology language overview. Technical report, W3C recommendation.

E. Montiel-Ponsoda, G.A. de Cea, A. Gómez-Pérez, and W. Peters. 2008. Modelling multilinguality in ontologies. In *Proceedings of the 21st International Conference on Computational Linguistics (COLING)*, pages 67–70.

W. Peters, E. Montiel-Ponsoda, G. Aguado de Cea, and A. Gómez-Pérez. 2007. Localizing ontologies in owl.

N. Smith, S. Hoffmann, and P. Rayson. 2008. Corpus tools and methods, today and tomorrow: Incorporating linguists manual annotations. *Literary and Linguistic Computing*, 23(2):163–180.

K. Toutanova and C.D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 conference on Empirical methods in natural language processing*, pages 63–70.

R. Trillo, J. Gracia, M. Espinoza, and E. Mena. 2007. Discovering the semantics of user keywords. *Journal of Universal Computer Science*, 13(12):1908–1935.

C. Unger and P. Cimiano. 2011. Pythia: Compositional meaning construction for ontology-based question answering on the semantic web. In *Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB 2011)*, pages 153–160.

D. Wu and P. Fung. 1994. Improving chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 180–181.

L. Zhou. 2007. Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3):241–252.

PromONTotion: Creating an Advertisement Thesaurus By Semantically Annotating Ad Videos Through Collaborative Gaming

Katia Lida Kermanidis[1], Emmanouil Maragkoudakis[2]

[1]Department of Informatics
Ionian University
7 Tsirigoti Square, 49100 Corfu, Greece

[2]Department of Information and Communication Systems Engineering
University of the Aegean
83200 Karlovasi, Samos, Greece

kerman@ionio.gr, mmarag@aegean.gr

Abstract

The present work describes the plan of PromONTotion, a ready to launch research project that aims at creating a semantic thesaurus of the advertising domain. The resource will be developed collaboratively using crowdsourcing. A web-based game, entertaining enough to keep the player's interest active for a long time, will be designed for the collaborative semantic annotation of the content of ad videos. The inserted terms will populate the thesaurus, a hierarchical structure formed by concepts, concept attributes and semantic relations among them. Advertisers will access the thesaurus through a friendly interface, which will allow them to have full access to the capabilities of the resource. The ad videos, the terminology, statistical information regarding co occurrence of concepts and attributes, statistical information regarding the impact the ads had on the annotators-players will be available to the advertiser for supporting him in the creative process of designing a new ad campaign.

1. Introduction

The present work describes the plan of a research proposal, PromONTotion, accepted for funding by the Greek National Strategic Reference Framework (NSRF) 2007-2013. The Department of Marketing of the Technological Institute of Thessaloniki, the Department of Informatics of the Ionian University and the Department of Information and Communication Systems Engineering of the University of the Aegean will form the project's research team. The project is planned to start in March 2012 and be completed by the end of 2014.

It should be noted at this point that the aim of the present paper is to sketch the outline of the project, as it entails several interesting research challenges related to the collaborative annotation of digital content for the creation of a semantic resource, i.e. a terminological thesaurus. As the project has not yet started, most technical aspects have not yet been investigated in depth, and several methodological issues still need to be clarified. However, despite the unavailability of this technical information and of any kind of results at this point, the highly interdisciplinary nature of the project and the innovative combination of tools, methodologies and application scenarios comprising it, make it, to the authors' opinion, worth writing about, even at this point.

2. Collaborative Annotation

Nowadays the research flow has shifted towards crowdsourcing for annotating data, especially data available on the web. Combining the effort of the public to label digital content provides an intriguing solution to the problem of annotation. The data to be annotated may be text, images, audio or video.

Several toolkits exist for the collaborative annotation of text. A solid study of the most important ones has been conducted by Wang et al. (2010). Approaches have also proposed gaming media (Chamberlain et al., 2008) to ensure the attraction of the annotator's interest, by making the annotation process as entertaining as possible.

Regarding the annotation of images, several collaborative annotation tools exist, like Catmaid (http://fly.mpi-cbg.de/~saalfeld/catmaid/), Flickr (www.flickr.com), Riya (www.riya.com), ImageLabeler (*http://images.google.com/imagelabeler)* and Imagenotion (Walter and Nagypal, 2007). ImageLabeler presents the same image to two players, and rewards the player who manages to annotate it with more semantically related terms in a given time frame. Von Ahn (2006) recognized that the high level of game popularity may be taken advantage of, and channelled towards other, more "serious", applications, instead of only pure entertainment. Imagenotion introduces the task of ontology maturing via assigning terms to images, and determining terms that are semantically related.

Tools for providing semantic descriptors for video content are also available, like VideoAnnex (http://mp7.watson.ibm.com/VideoAnnEx) and YouTube Collaborative Annotations (http://youtube-global.blogspot.com/2009/02/introducing-collaborative-annotations.html). Siorpaes and Hepp (2008) developed a more sophisticated game for the annotation of video content and Wikipedia terms.

PromONTotion will focus on the semantic annotation of video content, and, more specifically, advertisement videos available on the web. Annotations will be collected through a web-based game. The success of PromONTotion relies heavily on the plethora of provided annotations. The entertaining, engaging, and sometimes even addictive nature of videogames is the reason for choosing a videogame as the annotation tool, as it constitutes a promising option for gathering large amounts of annotations.

While gaming approaches to annotating text have been proposed in previous work, as mentioned earlier, the nature of the textual data has not allowed for the design of genuinely entertaining gaming software. The annotation of ad videos, however, inspires the design of software that can keep the player's interest and engagement level active for a very long time. Furthermore, the semantic annotations will form a semantic thesaurus for the advertising domain. Unlike the Imagenotion ontology, the thesaurus aimed at by PromONTotion is comprised of a backbone of a more elaborate set of concepts and relations, as well as statistical information regarding the terms inserted by players to populate the backbone. Finally, the thesaurus will be accessible to advertising experts as a support tool for creating a novel ad campaign.

3. PromONTotion

Creative advertising is governed nowadays by significant budget allocations and large investments. Several studies have been published regarding the impact of advertising (Amos, Holmes and Strutton, 2008; Aitken, Gray and Lawson, 2008), as well as creativity in advertising (Hill and Johnson, 2004). A number of creativity support tools have been proposed, that usually focus on forcing upon the advertiser a certain restricted way of thinking, using creativity templates (Goldenberg et al., 1999). Other support tools focus on decision making regarding the communication of the ad (Burke et al., 1990). Opas (2008) presents a detailed overview of advertising support tools.

The primary goal of PromONTotion is to develop a semantic thesaurus in the advertising domain that will function as a support tool for advertisers. The thesaurus will be created indirectly, through the use of a video game accessible to anyone. While playing, the user will indirectly annotate the content of available ad videos (there are over 300.000 ad videos available online (www.youtube.com)) and give his/her opinion on the impact of the ad. The semantic information will be organized in an ontological structure, which will be made usable to professionals in the advertising domain through a user-friendly interface. Advertisers will be able to see the content of old ads for related products, and thereby come up with new ideas, gain insight regarding the impact of previous campaigns from the players' evaluation, look for screenshots of videos using intelligent search, based not only on keywords, but on concepts. The innovative generic nature of the tool will allow it to be flexible, scalable, adjustable to the end user's needs. Most importantly, unlike creativity templates, the generic nature does not impose any sort of 'mold' or template to the creative advertiser's way of thinking.

The proposal faces a number of significant research challenges such as the game design, which needs to be attractive and captivating, in order to keep the players' interest at a high level for a long time. The ontology is an innovation by itself; its content (concepts, features, relations), coverage and representation are very interesting research issues. Mapping the output of the game, i.e. the annotations, to the ontological structure is a very intriguing task, as well as the end-user interface that needs to be friendly and make full use of the resource capabilities. The interdisciplinary nature of the proposed idea is challenging, as the areas of video game design, ontology engineering, human-computer interaction and advertising are linked together to produce an advertising support tool.

4. Designing the Advertising Thesaurus

Advertising experts will give their input on the concepts, categories, relations that are relevant to an advertising campaign. The parameters that determine the message to the consumer (e.g. tag lines), the ad content, as well as its artistic features play a significant role in designing the semantic ontology. The selected concepts describe the content of the video, which will include

- the characters involved (gender, roles, laymen, celebrities)
- the key objects involved
- the scenario or story taking place (the plot, the story, if one exists)
- the environment (where it takes place: indoors, outdoors etc)
- the ad genre (realistic, science fiction, animation etc.)
- the soundtrack of the ad
- the photography (the scenery, background,)
- the linguistic features of the ad (use of analogs, taglines, metaphors, paraphrasing etc.).

The aforementioned concepts will then be organized in a hierarchical structure, i.e. ontology, by ontology experts. They will be enriched with features that characterize them, as well as semantic relations that link them together. An example of such a structure is shown in Figure 1. For the development of the ontology, an ontology editor will be employed, like Protégé (http://protege.stanford.edu) or OntoEdit (www.ontoknowledge.org/tools/ontoedit.shtml). It is important for the ontology to be scalable, so it can constantly be enriched and updated.

5. Populating the Ontology

Based on the proposed ontology, a series of quiz questions will be designed, forming the core of the game. The questions' aim will be to acquire information that will populate the ontology. They will be clear, concise, requiring short answers so as to avoid confusing and tiring the player. Examples of questions are:

- Is there a dominating human presence in the ad? male/female/both/neither/irrelevant
- Does the ad remind you of a familiar movie script? If

yes, which one?
- Where does the main part of the ad take place? Indoors/City/Car/Country/Other

Each of these questions will correspond to a semantic relation or a concept attribute in the ontology backbone. Where possible, the multiple choice answers will be provided through visual and graphical means. Clicking or dragging and dropping of a graphical answer corresponds to inserting a term into the ontology. The suitability of Protégé or OntoEdit for taking the input terms and exporting them to the thesaurus will need to be determined. In every case, the editing platform needs to be as invisible to the player as possible, as these platforms are usually unfriendly and not usable for non ontology experts (Walter and Nagypal, 2007).

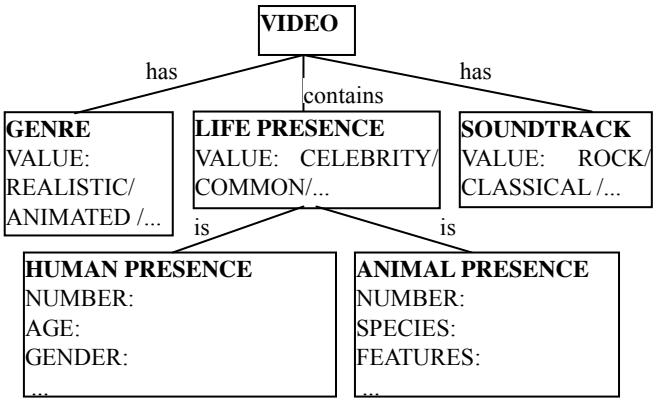

Figure 1: Example of a part of an advertising ontology

Apart from these 'objective' questions, a set of questions will focus on collecting the players' opinion regarding the ad's impact. For example questions like

- Does in your opinion the ad attack a competitive product? If yes, which one?
- What kind of impact does the advertisement have on you? I liked it/Indifferent/Boring/Influential

address the player's personal sentiment regarding the ad. There is a large number of available questionnaires for the evaluation of advertising campaigns (e.g. www.surveyshare.com/templates/televisionadvertisement evaluation.html and www-sea.questionpro.com/akira/showSurveyLibrary.do?surveyID=119&mode=1), based on which this type of questions may be formed.

A major asset of the developed thesaurus is the inclusion of statistical data, e.g. in how many ads for cleaning products there is a housewife in the leading role etc. To this end, co occurrence counts between concepts, attributes and terms will be extracted and statistical tests will be performed to investigate the significance of the co occurrences.

Data mining and machine learning techniques will be employed for post-processing the annotation data. The concepts and attributes described earlier will enable the transformation of the annotation set of a given advertisement into a learning feature-vector. The vectors will, in turn, enable feature selection, that will reveal the significance of the selected concepts on advertisement design (and indirectly evaluate the ontology design), as well as learning correlations between ad content choices, ad products and consumer impact. The extracted, mined, knowledge will reveal very interesting and previously unknown information regarding the parameters that directly or indirectly affect ad design and play a role on the consumers' sentiment and how the latter may be influenced.

6. Designing the Game

The game needed for the present work needs to be entertaining, interesting, not tiring and easy to play. Within the game, the player will be able to watch (part of) an advertisement video and face challenges regarding the ad. It will be designed for one, as well as multiple players, that compete against each other and against time. For each correct answer, the player gets points. The greater the challenge, the more points it will offer, if addressed correctly.

The correctness of the provided annotations will be evaluated through collaborative testing, i.e. if the answers of two or more players to the same ad question coincide, the term-answer is considered correct and is inserted into the thesaurus. IP address monitoring, random checking of sample answers and comparing them against the correct ones, keeping a blacklist of players with an untrustworthy annotating history, will allow the detection of cheating and irresponsible annotations.

The game will include multiple levels of difficulty (higher level indicates more fine-grained semantic distinctions). The set of questions will be grouped into levels of difficulty. At each level the player will be presented with a series of questions for that level, and he may address them all and move to the next level automatically, or skip questions and move to the next level at will, or abandon the particular ad and move to another one, or exit the game altogether at any time. Incomplete game sessions (sessions that have not provided answers to all questions) are not as important as attracting as many players as possible. A wide impact (acceptance of the game) will compensate for the lack of completeness. At every level, questions will appear in random order, so as to keep the interest of the user active on the one hand, and to ensure a higher degree of completeness of the various thesaurus slots.

The game will be available online and accessible by anyone. Social media and popular networks will be exploited for its dissemination and diffusion. The game will be evaluated according to several aspects. Its

usability will be tested through interviews and questionnaires handed out to a group of players, in combination with the talk aloud protocol, used extensively for the evaluation of the usability of interfaces. Another evaluation aspect is whether the game does indeed manage to populate the proposed ontology.

7. Creating the Support Tool

Advertisers will have access to the thesaurus via a user-friendly interface, that will allow them to make full use of the ontology's capabilities. The advertiser will be able

- to have access to a rich library of video ads
- to search the videos by content, based on a query of keywords (e.g. a specific type of product)
- to retrieve statistical data regarding the ads, i.e. see the terms/concepts/attributes his search keyword co-occurs with most frequently
- have access to the consumers' evaluation on the advertisements' impact.

The support tool will be evaluated by a group of advertising experts, who will be handed a product and will be asked to create a hypothetical ad scenario for it. They will evaluate the support tool based on its usability, its completeness, its significance. A combination of evaluation approaches will be employed in order to record the opinion and the impressions of the end users. Questionnaires will be handed out, interviews will be conducted to detect the problems and weaknesses of the support tool. Problems in the tool usability will be identified by evaluating its usage in real time with the think-aloud protocol. A group of ontology experts will evaluate the created ontology based on international ontology evaluation standards for its coverage, classification ability etc.

8. Conclusion

PromONTotion aims for the design, implementation and evaluation of a novel support tool for creative advertisers, that will facilitate their brainstorming process with the help of a semantic thesaurus. The thesaurus will be constructed automatically by consumers through game playing. While playing, they will annotate ad videos, describe the ad content and artistic features, and evaluate the ad impact on themselves. This information (the set of terms, concepts and subjective opinions), as well as statistical co occurrence data regarding concepts, advertised products, and subjective impact, will be structured into a hierarchical ontology. A user-friendly interface will allow ad designers to make full use of the ontology's capabilities, and advertising experts will evaluate the tool's coverage, usability and significance.

9. Acknowledgements

This Project is funded by the National Strategic Reference Framework (NSRF) 2007-2013: ARCHIMEDES III – Enhancement of research groups in the Technological Education Institutes. The authors are thankful for all this support.

10. References

Aitken, R., Gray, B. and Lawson R. (2008). Advertising Effectiveness from a Consumer Perspective. *International Journal of Advertising*, 27 (2), pp. 279–297.

Amos, C., Holmes, G. and Strutton, D. (2008). Exploring the Relationship between Celebrity Endorser Effects and Advertising Effectiveness. *International Journal of Advertising*, 27 (2), pp. 209–234.

Burke, R., Rangaswamy, A., Wind, J. and Eliashberg, J. (1990). A Knowledge-based System for Advertising Design. *Marketing Science*, 9(3), pp. 212–229.

Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008) Phrase Detectives: A Web-based Collaborative Annotation Game. Proceedings of *I-Semantics*.

Ericsson, K., and Simon, H. (1993). *Protocol Analysis: Verbal Reports as Data* (2nd ed.). Boston: MIT Press.

Goldenberg, J., Mazursky, D. and Solomon, S. (1999). The Fundamental Templates of Quality Ads. *Marketing Science*, 18(3), pp. 333–351.

Hill, R. and Johnson, L. (2004). Understanding creative service: A Qualitative Study of the Advertising Problem Delineation, Communication and Response (apdcr) process. *International Journal of Advertising*, 23(3), pp. 285–308.

Opas, T. (2008). *An Investigation into the Development of a Creativity Support Tool for Advertising*. PhD Thesis. Auckland University of Technology.

Siorpaes, K. and Hepp, M. (2008). Games with a Purpose for the Semantic Web. IEEE Intelligent Systems, pp. 1541-1672.

von Ahn, L. (2006). Games with a Purpose. *IEEE Computer*, 39 (6), pp. 92-94.

Walter, A. and Nagypal G. (2007) IMAGENOTION – Collaborative Semantic Annotation of Images and Image Parts and Work Integrated Creation of Ontologies, Proceedings of the *1st Conference on Social Semantic Web* (CSSW), pp. 161-166, Springer, LNCS.

Wang, A., Hoang, C. D. V. and Kan, M. Y. (2010). Perspectives on Crowdsourcing Annotations for Natural Language Processing. Technical Report (TRB7/10). The National University of Singapore, School of Computing.

The Phrase Detective Multilingual Corpus, Release 0.1

Massimo Poesio,[†] Jon Chamberlain,[†] Udo Kruschwitz,[†] Livio Robaldo, Luca Ducceschi

[†]University of Essex, University of Torino, University of Utrecht

Abstract

The Phrase Detectives Game-With-A-Purpose for anaphoric annotation has been live since December 2008, collecting over 2.5 million judgments on the anaphoric expressions in texts in two languages (English and Italian) from around 9,000 players. In this paper we summarize our recent work on creating a corpus using these annotations.

1. Introduction

Phrase Detectives, an interactive online **game with a purpose** (von Ahn, 2006) for creating anaphorically annotated resources making use of a highly distributed population of contributers with different levels of expertise, is an illustration of a new approach for creating large-scale resources: exploiting collective intelligence. In this paper we briefly discuss the language resources side of the enterprise–i.e., how the corpus has been prepared for annotation, the coding scheme, the data being annotated, and the agreement on the annotation.

2. The Game

Phrase Detectives is a single-player game-with-a-purpose developed to collect data about anaphora and centered around the detective metaphor. The game architecture is articulated around a number of **tasks** and uses scoring, progression and a variety of other mechanisms to make the activity enjoyable. A mixture of incentives, from the personal (scoring, levels) to the social (competing for some players, participating in a worthwhile enterprise for others) to the financial (small prizes) are employed.

2.1. Game Design

In *Phrase Detectives* the player is a **detective** that goes about resolving **cases**–expressing judgments about the interpretation of markables–in the so-called **Name-the-Culprit** activity, and providing opinions about other detectives's judgment in the **Detectives Conference** activity. Both of these activities lead to point accumulation, which is the main objective of the players; in fact, as we will see below, validation (Detectives Conference) is the main scoring activity for players once they pass the training threshold.

Name-the-Culprit Name-the-Culprit is the primary activity dedicated to the labelling of data by players. The players are shown a window of text in which a markable is highlighted in orange, as shown in Figure 1 (on the left).[1] They have to decide, first of all, whether the markable is referring, a property, or non-referring. In case they decide the markable is referring, they then have to decide whether it introduces a new entity (i.e., whether it is discourse new), or whether it refers to an already mentioned entity–and in this case they have to locate the closest mention. Moving

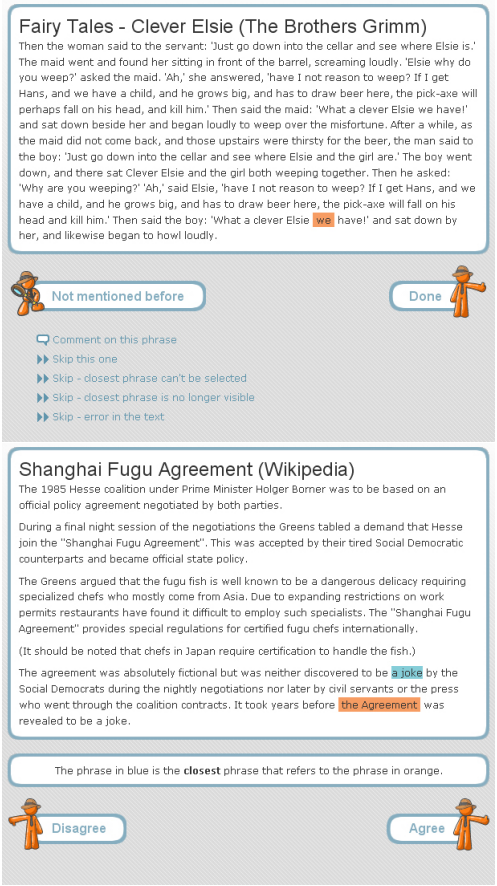

Figure 1: Screenshots of Annotation Mode (top) and Validation Mode (bottom)

the cursor over the text reveals the markables within a bordered box; to select a markable the player clicks on the bordered box and the markable becomes highlighted in blue.

Detectives Conference Every markable for which multiple interpretations have been proposed (the great majority, as discussed in Section 4.) must go through the validation process, **Validation Mode**–aka the **Detectives Conference** activity, displayed on the right side of Figure 1. In Detectives Conference players have to say whether they agree or disagree with an interpretation.

2.2. Other Points

The game-with-a-purpose approach to resource annotation was adopted not just to annotate large amounts of text, but also to collect a large number of judgments about each lin-

[1]These markables are automatically extracted from the text using the pipeline(s) discussed below.

guistic expression, which led to the deployment of a variety of mechanisms for quality control which try to reduce the amount of unusable data beyond those created by malicious users, from the level mechanism itself to validation to a number of tools for analyzing the behavior of players. More recently, a Facebook version was developed.

3. The Corpus

The ultimate goal of Phrase Detectives is to obtain very large anaphorically annotated corpora for the languages covered (currently, English and Italian).

3.1. Coding Scheme

The *Phrase Detectives* corpus is annotated according to the linguistically-oriented approach to anaphoric annotation that is currently prevalent, having been adopted in OntoNotes (Pradhan et al., 2007), our own ARRAU corpus (Poesio and Artstein, 2008) and in all the corpora used in the 2010 SEMEVAL anaphora evaluation (Recasens et al., 2010). In this type of annotation, all NPs are considered markables, and anaphoric relations between all types of entities are annotated, unlike the practice in the MUC and ACE corpora.[2] (E.g., in the *Phrase Detectives* corpora, coordinated NPs like *John and Mary* are also markables.)

Players can assign four types of interpretation (labels) to markables:

- `DN` (discourse-new): the markable refers to a newly introduced entity.
- `DO` (discourse-old): the markable refers to an already mentioned entity; the player has to specify the latest mention.
- `NR` (non-referring): the markable is non-referring (e.g. pleonastic *it*).
- `PR` (property attribute): the markable represents a property of a previously mentioned entity (e.g., *a teacher* in "He is a teacher").

3.2. Input / Output

The data handled by *Phrase Detectives* are stored in a relational database whose design for the part concerned with storing texts and their annotations is based on that of the University of Bielefeld's Serengeti system (Poesio et al., 2011). New texts are entered in the system through the Serengeti interface, that requires input in SGF format (Stührenberg and Goecke, 2008). The text must have been preprocessed to identify tokens, sentences, and noun phrases. The data are outputted in an extended version of the MAS-XML format (Kabadjov, 2007), designed to represent anaphoric information and to encode multiple interpretations. The extended version of MAS-XML, called PD-MAS-XML, can be used to export each interpretation assigned to each markable in the text.

[2] `http://projects.ldc.upenn.edu/ace/data/`

3.3. MAS-XML

The PD-MAS-XML format used to export *Phrase Detectives* data is a modified version of the Minimum Anaphoric Syntax (MAS-XML) format proposed in (Kabadjov, 2007). MAS-XML is a form of inline XML in which the basic information required to carry out resolution is marked, including

- sentences;
- words with their part-of-speech tags (for English, the Penn Treebank tagset is used);
- NPs (called Nominal Entities, `ne`), with their ID and the basic agreement features: gender (attribute `gen` for gold-standard info, `AAgen` for automatically extracted information), number (again two attributes are used, `num` and `AAnum`), and person (using the attributes `per` and `AAper`)
- NP modifiers and heads, using the elements `mod` and `nphead`

Anaphoric information is marked using separate `ante` elements, a structured representation inspired by the Text Encoding Initiative `link` elements and that makes it possible to specify multiple anaphoric relations for each markable (identity and association) and to mark ambiguity using multiple `anchor` elements (Poesio, 2004).

The MAS-XML file for each document that is exported contains the original text and markup (sentences, NPs and their features and constituents) automatically computed by the import pipeline, as well as the annotations produced by the players. To export the annotation information, the anchor mechanism from MAS-XML was replaced by a much more extensive format specifying for every player that expressed a judgment about a given markable the interpretation (`DN` for Discourse-New, `DO` for Discourse-Old, `NR` for Non-Referring, or `PR` for Property), any antecedents selected for `DO` and `PR` interpretations, the user ID, the user rating, the time it took to make the annotation, whether the decision is an agreement and in what mode the decision occurred (annotation or validation). Additionally players' comments are exported with the relevant markable and include the user ID, the type of comment and the text that was submitted; and so are skips. For instance the (real-life) interpretation of markable `ne14817`, which all players interpreted as `DN`, is as follows.

```
<PDante id="ne14817">
   <interpretation>
      <anchor type="DN" user_id="281" user_rating="75"
           annotation_time="2" agree="y" mode="a"/>
      <anchor type="DN" user_id="728" user_rating="58"
           annotation_time="2" agree="y" mode="a"/>
      <anchor type="DN" user_id="779" user_rating="77"
           annotation_time="5" agree="y" mode="a"/>
      <anchor type="DN" user_id="281" user_rating="75"
           annotation_time="1" agree="y" mode="a"/>
      <anchor type="DN" user_id="18" user_rating="77"
           annotation_time="5" agree="y" mode="a"/>
      <anchor type="DN" user_id="1293" user_rating="64"
           annotation_time="15" agree="y" mode="a"/>
      <anchor type="DN" user_id="1364" user_rating="59"
           annotation_time="4" agree="y" mode="a"/>
      <anchor type="DN" user_id="163" user_rating="80"
           annotation_time="2" agree="y" mode="a"/>
      <anchor type="DN" user_id="1659" user_rating="92"
           annotation_time="9" agree="y" mode="a"/>
   </interpretation>
   <skip total="0"/>
</PDante>
```

Documents can be exported from *Phrase Detectives* in MAS-XML format either when they are complete (i.e when all the markables have been annotated sufficiently according to the game configuration) or when they are partially complete. For the purposes of testing only complete documents have been exported.

3.4. Preprocessing

Adding texts in a new language to *Phrase Detectives* requires developing a **pipeline** to convert documents into SGF format importable in the database. Two such pipelines have been developed so far.

The English Pipeline The English *Phrase Detectives* pipeline converting raw text to SGF was developed by combining existing tools (OpenNLP tokenizer and sentence splitter, Berkeley Parser) with *ad-hoc* modules for correcting the output of such tools in the case of frequent errors.

The Italian Pipeline In order to use *Phrase Detectives* to annotate Italian data, a new pipeline (Robaldo et al., 2011) was developed using the TULE parser (Lesmo and Lombardo, 2002). The parser processed the raw text directly with Italian texts so no pre-processing is needed.

3.5. The English and Italian Corpora

As our ultimate goal is to produce a freely distributable corpus, the texts of the English and Italian corpus are from collections not subject to copyright restrictions.

English The English texts come from three main domains:

- Wikipedia articles selected from the 'Featured Articles' page[3] and the page of 'Unusual Articles'[4];
- narrative text from Project Gutenberg[5] including in particular a number of tales (e.g., Aesop's Fables, Grimm's Fairy Tales, Beatrix Potter's tales), and more advanced narratives such as several Sherlock Holmes short stories by A. Conan-Doyle, *Alice in Wonderland*, and several short stories by Charles Dickens.
- dialogue texts from Textfile.[6]

The ultimate objective is to annotate over 100 million words, and several millions words of text have already been converted, but in part because the accuracy of the present pipeline is not considered high enough, at present only around a million words have been actually uploaded in the English version of *Phrase Detectives*–to be precise, 1,206,597 words from 839 documents.

Italian The same criteria concerning distribution were used for the texts in the Italian version of the game; an additional criterion has been the kind of linguistic phenomena that they are likely to include. The sources are the Italian version of Wikipedia and two novels by Wu Ming (CC licensed).

The texts from Wikipedia belong to two specific sub-genres (plots and biographies) which are likely to contain a dense net of antecedents. The first kind displays a significant number of pronominal anaphors, while the second might display examples of lexical noun phrase anaphora (e.g., "the Queen" and "her Majesty.") In addition to the mentioned sub-genres other uncategorized texts have been chosen in order to provide a comparison with the English version of the game ("Chess Boxing" and "Diet Coke and Mentos Explosion" are in both corpora).

The novels have been selected to test if the narrative style has an influence on the performance of the parser and of the players. This variety is more likely to display all the pronouns of the language, particularly 1st and 2nd person in reported speech, which are less likely to appear in Wikipedia articles.

The Italian corpus for *Phrase Detectives* currently contains 30 texts, for a total of 11,373 words.

4. Results So Far

4.1. A Quantitative Assessment

Since the first release of the game in December 2008 to January 2012 just over 10,000 players have registered (10,250 as this paper is completed), 2,000 of which went beyond the initial training phase. 665 of these players are using the Facebook version.

445 documents have been fully annotated, for a total completed corpus of 181,000 words, 15% of the total size of the collection currently uploaded for annotation in the game (1.2M words). This is comparable in size to the ACE2 corpus of anaphoric information (BNews + Npaper + Nwire),[7], which was the standard for evaluation of anaphora resolution systems until 2007/08 and still widely used. The size of the completed corpus does not properly reflect, however, the amount of data we have collected, as the case allocation strategy adopted in the game privileges variety over completion rate; as a result, almost all the 841 documents in the corpus have already been partially annotated. This is reflected, e.g., in the fact that 84280 of the 392,120 markables in the active documents (21%) have already been annotated. This is already almost twice the total number of markables in the entire OntoNotes 3.0 corpus,[8] which contains 1 million tokens, but only 45,000 markables.

4.2. Agreement on Annotations

In order to check the extent to which the annotations produced by the game corresponded to the annotations produced by experts, we randomly selected five completed documents from the Wikipedia corpus containing 154 markables. Each document was manually annotated by two experts (called Expert 1 and Expert 2 in the rest of this discussion) operating separately; we then compared the annotations produced by the experts with the most highly

[3] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

[4] http://en.wikipedia.org/wiki/Wikipedia:Unusual_articles

[5] http://www.gutenberg.org/

[6] http://www.textfiles.com/

[7] http://projects.ldc.upenn.edu/ace/data/

[8] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T24

ranked interpretations produced by the players (henceforth, the **game interpretation**), and with each other.

Overall, agreement between experts on the types is very high although not complete: 94%, for a chance-adjusted κ value (Artstein and Poesio, 2008) of $\kappa = .87$, which is extremely good. This value can be seen as an upper boundary on what we might get out of the game. Agreement between each of the experts and the majority interpretation of the game is also good: we found 84.5% percentage agreement between Expert 1 and the game ($\kappa = .71$) and 83.9% agreement between Expert 2 and the game ($\kappa = .7$). In other words, in about 84% of all cases the interpretation specified by the majority vote of non-experts was identical to the one assigned by an expert. These values are comparable to those obtained when comparing an expert with the 'normally trained' annotators (usually students) that are typically used to create medium-quality resources.

4.3. Ambiguity in the Corpus

We are in the process of analyzing the judgments accumulated so far in preparation for a paper on anaphora through the lens of *Phrase Detectives*, and some interesting results already came up, in particular about the notion of coreference (e.g., in many mysteries, the whole point of the story is that the identity of a character–the culprit, or some shady figure– is only discovered at the end). We will not enter into this discussion here, but one preliminary statistic is worth reporting given the motivating role that studying anaphoric ambiguity has had in the design of the game. In January 2011 there were 63009 completely annotated markables. Of these, 23479 (37.3%) had exactly one interpretation (i.e., the first eight players to be presented with that markable all chose the same interpretation). Of these, 23,138 were DN, 322 DO, and 19 NR. A further 13,772 markables (21%) had only 1 interpretation with a score greater than 0. Again, the majority of these (9,194) were DN; 4,391 were DO, and NR 175.

5. Discussion

Phrase Detectives was one of the very first GWAP applied to resource creation for HLT and in quantitative terms has been the most successful, collecting over 2.5 million judgments from over 10,000 players. Annotation is still going strong and we expect it to continue for the immediate future; our hope is to complete at least the annotation of the initial 1.2M corpus of documents. In order to annotate more data, a higher-quality preprocessing pipeline for English will be required.

Among the lessons we learned, the first and most obvious is that GWAP can be used for HLT resource creation. However researchers will need to consider with great care whether in fact this approach is appropriate for their task and their data. If only a small amount of data is required (100,000 words or less), and / or the data are not very interesting, it may be best to use crowdsourcing instead. If the GWAP approach is chosen, a constant effort of promotion will be required to make the game stand out among the thousands of other games (serious or not)–but offering small prizes proved very effective.

Concerning the architecture of the game, the main lesson we learned is that validation is essential and very effective for quality control. Keeping around all interpretations also proved the right choice. Last but not least, embedding the game in Facebook has proven very effective not so much as a new way of reaching players but to know better who your players are.

Next steps include developing methods for cleaning up the data and for using the data to train anaphoric models.

6. Acknowledgements

The initial funding for *Phrase Detectives* (2007/09) came from EPSRC project AnaWiki, EP/F00575X/1.

7. References

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

M. A. Kabadjov. 2007. *Task-oriented evaluation of anaphora resolution*. Ph.D. thesis, University of Essex.

L. Lesmo and V. Lombardo. 2002. Transformed subcategorization frames in chunk parsing. In *Proc. of LREC*, pages 512–519, Las Palmas.

M. Poesio and R. Artstein. 2008. Anaphoric annotation in the arrau corpus. In *Proc. of LREC*, Marrakesh, May.

M. Poesio, N. Diewald, M. Stührenberg, J. Chamberlain, D. Jettka, D. Goecke, and U. Kruschwitz. 2011. Markup infrastructure for the anaphoric bank: Supporting web collaboration. In A.Mehler, K.-U. Kühnberger, H. Lobin, H. Lüngen, A. Storrer, and A. Witt, editors, *Modeling, Learning, and Processing of Text Technological Data Structures*, Springer, pages 175–195.

M. Poesio. 2004. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL*.

S. S. Pradhan, L. Ramshaw, R. Weischedel, J. MacBride, and L. Micciulla. 2007. Unrestricted coreference: Indentifying entities and events in ontonotes. In *Proc. ICSC*, Irvine, CA.

M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proc. SEMEVAL*, Uppsala.

L. Robaldo, M. Poesio, L. Ducceschi, J. Chamberlain, and U. Kruschwitz. 2011. Italian anaphoric annotation with the Phrase Detectives game-with-a-purpose. In *Proc. of AIIA*, Lecture Notes in Artificial Intelligence, pages 407–412, Berlin. Springer.

M. Stührenberg and D. Goecke. 2008. SGF – An integrated model for multiple annotations and its application in a linguistic domain. In *Balisage: The Markup Conference*, Montreal, Kanada.

L. von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Towards a Linguistic Linked Open Data cloud: Linking the MASC

Christian Chiarcos

Information Science Institute, University of Southern California
chiarcos@daad-alumni.de

Abstract

I describe benefits of modeling linguistic resources as Linked Data, i.e., using RDF, publishing them under an open licence, and creating links between them. Further, an overview over currently on-going community efforts to create a Linked Open Data (sub-)cloud of linguistic resources will be given. Both aspects are illustrated for the MASC corpus.

1. Overview

Nowadays, computational linguistics, Natural Language Processing and Information Technology are confronted with an immense – and steadily growing – wealth of linguistic resources accumulated in more than half a century of computational linguistics (Dostert, 1955), of empirical, corpus-based study of language (Francis and Kučera, 1964), and of computational lexicography (Morris, 1969).

To make these resources available to the different communities interested in linguistic resources, and to facilitate their use, however, a number of technological challenges are to be addressed. One fundamental problem is the interoperability of existing language resources, a problem actively addressed by the community since the late 1980s (Text Encoding Initiative, 1990), but still a problem that is partially solved at best (Ide and Pustejovsky, 2010). A closely related challenge is information integration, i.e., how heterogeneous information from different sources can be retrieved and combined in an efficient way. To address both problems, the linguistic and NLP communities are developing generic standards for different types of linguistic resources, including the Lexical Markup Framework (Francopoulo et al., 2006, LMF) for lexical-semantic resources and the Graph Annotation Framework (Ide and Suderman, 2007, GrAF) for annotated corpora, both maintained by the ISO TC37/SC4.

Outside the linguistic community, similar problems have been addressed, for example, in the discussion of meta data for the world wide web. The formalisms proposed there eventually converged into the Resource Description Framework (Klyne et al., 2004, RDF, W3C recommendation 1999). RDF provides very generic data structures (labeled directed multi-graphs), that were applicable to a broader band-width of problems than originally anticipated. Hence, RDF was readily adopted in other domains, and employed for different tasks. Nowadays, RDF represents the state of the art of knowledge representation in many scientific disciplines, and eventually it became one of the fundamental elements of the Semantic Web. Because of its genericity, its further development was (and is) supported by a large and interdisciplinary community of developers and users, from academics as well as from industry. As a result, a rich technological ecosystem evolved, which includes different representation formats with varying degrees of compactness and readability (e.g., RDF/XML, RDF/Turtle, RDF/HDT),[1] specialized sub-languages for different tasks (e.g., RDFS for hierarchical structures, SKOS for semi-structured terminology bases, and OWL/DL for formally defined ontologies),[2] parsers, validators and (for OWL/DL) reasoners, several data bases (RDF triple stores) and query languages.

The potential of RDF for representing linguistic resources has long been recognized, in particular for lexical-semantic resources, where RDF can be employed to achieve interoperability between lexical resources with Semantic Web technologies (Gangemi et al., 2003), but also for linguistic corpora, where RDF technologies can be used to process, to store and to query multi-layer corpora (Burchardt et al., 2008).

In the talk, I briefly describe advantages of RDF for modeling linguistic resources, and in particular, linguistic corpora, using the Manually Annotated Sub-Corpus of American English (Ide et al., 2008, MASC) as an example. Aside from emphasizing the availability of infrastructures for efficiently storing and querying RDF data, I focus on two aspects, **interoperability** between different types of language resources, and **integration of information**. RDF extends resource-type specific formalisms like GrAF or LMF in that it establishes interoperability and information integration not only *for* annotated corpora or lexical-semantic resources, but also *between* both types of resources. Certainly, the LMF and the GrAF data model will guide the future development of standards for linguistics, but adding RDF as another possible serialization of these data models (along with classical XML linearizations) may open up the possibility to benefit from RDF infrastructures for specific tasks such as data integration, storing and querying. For the future of GrAF, this may mean that it evolves in a similar way as the LMF, i.e., that it gains a status as meta model for which multiple, but convertible linearizations in different formats are provided.

Interoperability of RDF data and information integration involve the **Linked (Open) Data paradigm** (Berners-Lee, 2006) that postulates four rules for the publication and representation of web resources: (1) Referred entities should be designated by using URIs, (2) these URIs should be resolvable over HTTP, (3) data should be represented by

[1] `http://www.w3.org/TR/rdf-syntax-grammar`, `http://www.w3.org/TR/turtle`, `http://www.w3.org/Submission/HDT`

[2] `http://www.w3.org/TR/rdf-schema`, `http://www.w3.org/TR/skos-reference`, `http://www.w3.org/TR/owl2-overview`

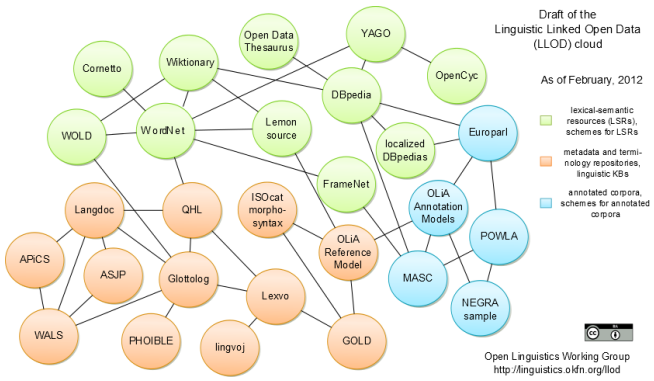

Figure 1: The Linguistic Linked Open Data (LLOD) diagram, draft version.

means of standards (such as RDF), (4) and a resource should include links to other resources. These rules establish information integration in that they require that entities can be addressed in a globally unambiguous way (1), that they can be accessed (2) and interpreted (3), and that entities that are associated on a conceptual level are also physically associated with each other (4).

The concept of Linked Data is closely coupled with the idea of **openness** (otherwise, the linking is only reproducible under certain conditions), and the definition of Linked Open Data has been extended with a 5 star rating system for data on the web. The first star is achieved by publishing data on the web (in any format) under an open license, the second, third and fourth star require machine-readable data, a non-proprietary format, and using standards like RDF, respectively. The fifth star is achieved by linking the data to other people's data to provide context.

If (linguistic) resources are published in accordance with these rules, it is possible to follow links between existing resources to find other, related data and exploit network effects. Following this insight, recent community efforts converge towards the development of a Linked Open Data (sub-)cloud of linguistic resources, the Linguistic Linked Open Data (LLOD) cloud, under the umbrella of the **Open Linguistics Working Group (OWLG)** of the Open Knowledge Foundation (Chiarcos et al., 2012). The OWLG is a multi-disciplinary network of researchers aiming to promote the idea of openness for linguistic resources, and dedicated to discussing and documenting the problems and benefits arising from open data in linguistics. It covers diverse disciplines, including language documentation, typology, computational linguistics, and information technology, just to name a few, and this diversity is also reflected in the current draft of the OWLG as illustrated in Fig. 1, which comprises general-purpose semantic knowledge bases (e.g., DBpedia), lexical resources (e.g., WordNet), annotated corpora (e.g., MASC), terminology repositories (e.g., an OWL linearization of the morphosyntactic profile of ISOcat), bibliographical data bases (e.g., Langdoc), and typological data bases (e.g., the World Atlas of Syntactic Structures, WALS).

I describe the integration of MASC in the LLOD cloud, and concrete use cases for the respective links.

2. References

Tim Berners-Lee. 2006. Design issues: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`. includes a 2010 addendum about linked *open* data.

Aljoscha Burchardt, Sebastian Padó, Dennis Spohr, et al. 2008. Formalising multi-layer corpora in OWL/DL – Lexicon modelling, querying and consistency control. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, Hyderabad, India, Jan.

Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, et al. 2012. The Open Linguistics Working Group. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May.

Leon E. Dostert. 1955. The Georgetown-IBM experiment. In William N. Locke and Andrew D. Booth, editors, *Machine Translation of Languages*, pages 124–135. John Wiley & Sons, New York.

W. Nelson Francis and Henry Kučera. 1964. Brown Corpus manual. Manual of information to accompany A standard corpus of present-day edited American English, for use with digital computers. Technical report, Brown University, Providence, Rhode Island.

Gil Francopoulo, Monte George, Nicoletta Calzolari, et al. 2006. Lexical Markup Framework (LMF). In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 233–236, Genoa, Italy.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003. Sweetening WordNet with DOLCE. *AI magazine*, 24(3):13.

Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proceedings of the 2nd International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, China.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the 1st Linguistic Annotation Workshop (LAW 2007)*, pages 1–8, Prague, Czech Republic.

Nancy Ide, Collin F. Baker, Christiane Fellbaum, et al. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Graham Klyne, Jeremy J. Carroll, and Brian McBride. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C Recommendation.

William Morris, editor. 1969. *The American Heritage Dictionary of the English Language*. Houghton Mifflin, New York.

Text Encoding Initiative. 1990. TEI P1 guidelines for the encoding and interchange of machine readable texts. Technical report, Text Encoding Initiative. Draft Version 1.1 1, `http://www.tei-c.org/Vault/Vault-GL.html`.