Special Topics in Computer Science

NLP in a Nutshell

CS492B Spring Semester 2009

Jong C. Park

Computer Science Department

Korea Advanced Institute of Science and Technology

ENCODING, ENTROPY, AND ANNOTATION SCHEMES

Encoding Texts

- American Standard Code for Information Interchange (ASCII)
  - Established originally for English
  - The adaptation of ASCII to other languages led to clunky evolutions and many variants.
- Unicode
  - A universal scheme compatible with ASCII and intended to cover all the scripts of the world

Character Sets

- Representing Characters
  - Codes are used to represent characters.
    - The Baudot code uses five bits, representing $2^5 = 32$ characters.
  - ASCII has 0 to 127 code points and is only for English.
  - The ISO Latin 1 character set (ISO-8859-1) extends it to 256 code points.
    - It can be used for most Western European languages but forgot many characters, like the French Œ, œ, the German quote „, or the Dutch *IJ, ij.*
  - ISO Latin 9 (ISO-8859-15) updated Latin 1.
    - It restored forgotten French and Finnish characters , and added the euro currency sign, €.

Character Sets

- Unicode
  - The Unicode consortium has produced character tables of most alphabets and scripts of European, Asian, African, and Near Eastern languages, and assigned numeric values to the characters.
  - It started with a 16-bit code that could represent up to 65,000 characters. It has subsequently been extended to 32 bits.

Character Sets

- Unicode
  - The Universal Character Set (UCS) is the standardized name of the Unicode character representation.
    - The 2-octet code (UCS-2) is called the Basic Multilingual Plane (BMP). All common characters fit on 16 bits, with the exception of some Chinese ideograms.
    - The 4-octet code (UCS-4) can represent more than a million characters. They cover all the UCS-2 characters and rare characters: historic scripts, some mathematical symbols, private characters, etc.

Character Sets

- Unicode
  - Examples
    - U+0041 LATIN CAPITAL LETTER A
    - U+0042 LATIN CAPITAL LETTER B
    - …
    - U+00CA LATIN CAPITAL LETTER E WITH CIRCUMFLEX
    - U+00C5 LATIN CAPITAL LETTER A WITH RING ABOVE
    - U+0391 GREEK CAPITAL LETTER ALPHA
    - …
  - The resulting graphical symbol is called a grapheme.

Character Sets

- The Unicode Encoding Schemes
  - Unicode offers three different encoding schemes: UTF-8, UTF-16, and UTF-32.
    - The UTF (Unicode Transformer Format) schemes encode the same data by units of 8, 16, or 32-bits and can be converted from one to another without loss.
  - UTF-16 was to be the standard encoding scheme.
    - It uses fixed units of 16 bits, or 2 bytes.
    - *FÊTE* ***0046 00CA 0054 0045***
    - ***FE^TE 0046 0045 0302 0054 0045***
  - UTF-8 is a variable length encoding.
    - It maps the ASCII code characters U+0000 to U+007F to their byte values 00 to 7F.
    - All the other characters in the range U+007F to U+FFFF are encoded as a sequence of two or more bytes.

Locales and Word Order

- Presenting Time, Numerical Information, and Ordered Words
  - Depending on the language, dates, numbers, time is represented differently:
    - Numbers: 3.14 (English) or 3,14 (most European languages)
    - Time: February 23, 2003 (US English); 24 February, 2003 or February 24th, 2003 (England); 2/24/03 (US); 24/02/2003 (Britain); 2003/02/24 (Sweden)

Locales and Word Order

- The International Organization for Standardization (ISO) has standardized the identification of languages and communities under the name of locales.
  - Each locale uses a set of rules that defines the format of dates, times, numbers, currency, and how to sort – collate – strings of characters.
  - A locale is defined by three parameters: the language, the region, and the variant that corresponds to more specific conventions used by a restricted community.

Locales and Word Order

- A locale has features including the collation component that defines how to compare and order strings of characters.
  - Elementary sorting algorithms consider the ASCII or Unicode values with a predefined comparison operator such as the inequality predicate @</2 in Prolog.
- They determine the lexical order using the numerical ranking of the characters. The lexicographic ordering of words varies from language to language.
  - Collating strings:
    - Is Andersson before of after Åkesson?

Locales and Word Order

- The Unicode Collation Algorithm
  - Takes into account the different practices and cultures in lexical ordering.
  - It has three levels for Latin scripts:
    - The primary level considers differences between base characters, for instance between A and B.
    - If there are no differences at the first level, the secondary level considers the accents on the characters.
    - And finally, the third level considers the case differences between the characters.

Locales and Word Order

- The Unicode Collation Algorithm
  - These level features are general, but not universal.
  - Accents are a secondary difference in many languages.
    - But Swedish sorts accented letters as individual ones and hence sets a primary difference between A and Å or o and Ö.
    - First level: {a, A, á, Á, à, À, etc.} < {b, B} < {c, C, ć, Ć, ĉ, Ĉ, ç, Ç, etc.} < {e, E, é, É, è, È, ê, Ê, ë, Ë, etc.} < ...
    - Second level: {e, E} << {é, É} << {è, È} << {ê, Ê} << {ë, Ë}
    - Third level: {a} <<< {A}
  - The comparison at the second level is done from the left to the right of a word in English, the reverse in French.

Locales and Word Order

- Lexical order of words with accents

English	French
Péché	*pèche*
PÉCHÉ	*pêche*
pèche	*Pêche*
pêche	*Péché*
Pêche	*PÉCHÉ*
pêché	*pêché*
Pêché	*Pêché*
pécher	*pécher*
pêcher	*pêcher*

Markup Languages

- A brief background
  - Texts in corpora are often annotated using a markup language.
  - Annotation schemes used by word processors include LaTeX, RTF, etc.
  - XML, which resembles HTML, enjoys a growing popularity as a corpus annotation language.
  - XML is a coding framework: a language to define ways of structuring documents.

Markup Languages

- An Outline of XML
  - XML separates the definition of structure instructions from the content – the data.
  - Structure instructions are described in a Document Type Definition (DTD) that models a class of XML documents.
  - Document Type Definitions contains the specific tagsets to mark up texts.
  - A DTD lists the legal tags and their relationships with other tags.

Markup Languages

- DTD
  - DTD is composed of three kinds of components: elements, attributes, and entities.
  - Elements are the logical units of an XML document.

```
<!-- My first XML document -->
<book>
<title>Language Processing Cookbook</title>
<author>Pierre</author>
<text>Here comes the text!</text>
</book>
```

Markup Languages

- DTD
  - An element can have attributes, i.e. a set of properties.
  - A <title> element can have an alignment: flush left, right, or center, and a character style: underlined, bold, or italics.

    <title align="center" style="bold">
    Language Processing Cookbook
    </title>
  - Entities are data stored somewhere in a computer.
    - They can be accented characters, symbols, strings as well as text or image files.
    - An entity is referred to using the start delimiter '&' and the end delimiter ';' such as &EntityName;.
      - **Example: $lt; (less than)**

Markup Languages

- Writing a DTD
  - The description of the elements is enclosed between the delimiters **<!ELEMENT and >.**

    <!ELEMENT book (title, (author | editor)?, chapter+)>

    <!ELEMENT title (#PCDATA)>
  - Attributes are the possible properties of the elements.
  - Their description is enclosed between the delimiters **<!ATTLIST** and **>.**

    <!ATTLIST title

    style (underlined | bold | italics) "bold"

    align (left | center | right) "left">

Markup Languages

- Writing an XML Document

```
<?xml version="1.0"
    encoding="UTF-8"?>
<!DOCTYPE book [
<!ELEMENT book
    (title, (author | editor)?,
     chapter+)>
<!ELEMENT title (#PCDATA)>
...
]>
<book>
<title style="i">La Grande
    Cuisine</title>
<author style="b"> Egon
    Lenoir</author>
<chapter number="c1">
<subtitle>Introduction</subtitle
    >
<para>Let&apos;s start doing
    simple things:
    How to fry an egg.</para>
<para>First, take a fresh egg.
    Break it into… </para>
</chapter>
</book>
```

Markup Languages

- Parsing XML
  - SWI Prolog has a package to process XML files: http://www.swiprolog.org/packages/sgml2pl.html
  - It makes it very easy to load and parse an XML document.
  - The most useful predicate is:
    - **load_xml_file(+File, -ListOfContent)**
  - as in
    - **load_xml_file('MyBook.xml', Term), write(Term).**
  - The element predicate has the form:
    - **element(Name, ListAttributes, ListOfContent)**

Codes and Information Theory

- Entropy
- Huffman Encoding
- Cross Entropy
- Perplexity and Cross Perplexity

Entropy and Decision Trees

- Decision Trees
- Inducing Decision Trees Automatically

SOFTWARE REVIEW 1 - AN OVERVIEW OF LANGUAGE PROCESSING

Today's Topics

- NLP Software Collections
  - The German Institute for Artificial Intelligence Research
- Lexical & Corpus Resources
  - The Oxford Text Archive
  - The Linguistic Data Consortium of the University of Pennsylvania
  - The European Language Resources Association

The German Institute for Artificial Intelligence Research

- http://registry.dfki.de/

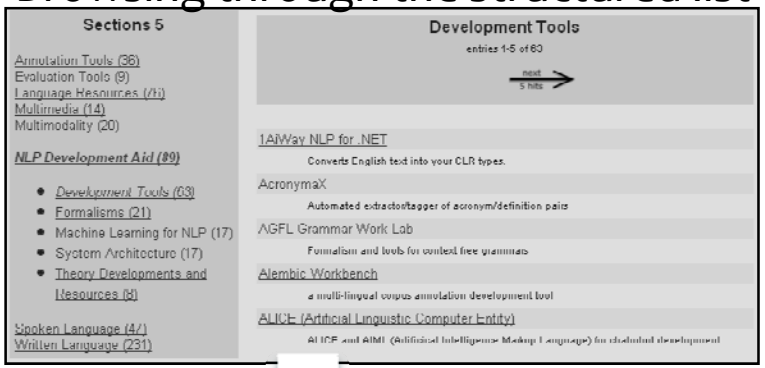
- A concise summary of the capabilities and sources of natural language processing (NLP) software

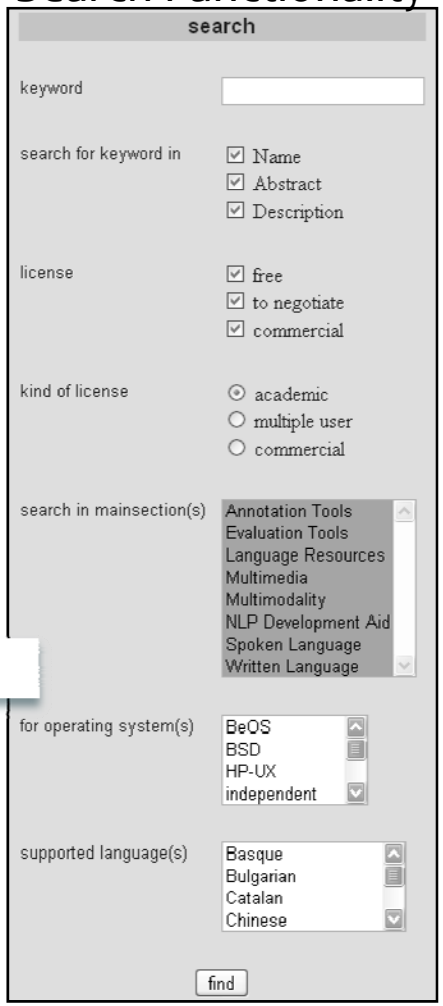
- Features
  - Browsing through the structured list
  - Finding products by submitting a query
  - Screening out submitted products

The German Institute for Artificial Intelligence Research

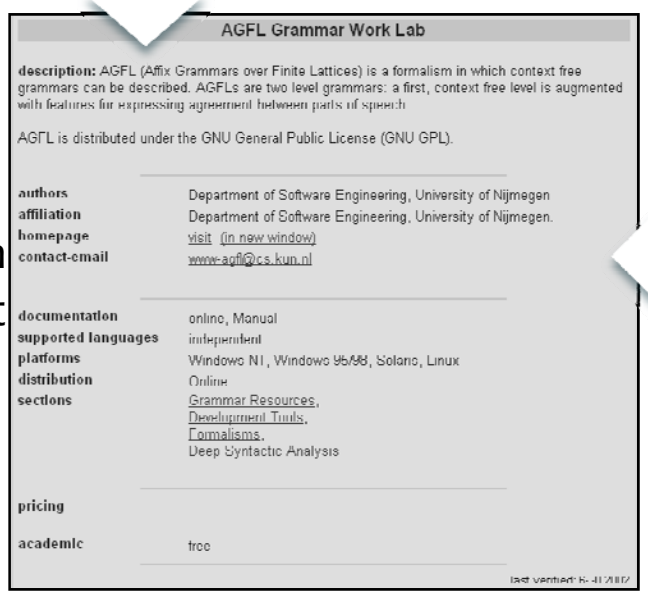
Browsing through the structured list

Search Functionality





Detailed Information on each product



The German Institute for Artificial Intelligence Research

- Pros
  - Easy to find what kind of software exists without knowledge through categories
  - Reliable quality of listed products which experts, the NLR team, review
- Cons
  - Hard to compare products with several factors because the list has only their descriptions
  - Many interactions needed to check all products within a category because of the fixed number of products per page

The Oxford Text Archive

- http://ota.ox.ac.uk

- A repository of digital literary and linguistic resources for research and teaching

- Features
  - Providing more than 1000 corpora
  - Screening out corpora

The Oxford Text Archive

Corpora list

ID	Availability	Title	Language	Author
2540	free	Speech, Thought and Writing Presentation Corpus (STWP)	English	Culpeper, Jonathon; Gemino, Elena; Short, Mick; Wynne, Martin
2539	restricted	British Academic Written English Corpus	English	Nesi, Hilary; Gardner, Sheena; Thompson, Paul; Wickens, Paul
2531	free	Thomason collection of Civil War tracts : copyflo set	English	Thomason, George, d. 1666
2530	restricted	Language convergence and grammatical borrowing database	English	
2529	restricted	The Workdiaries of Robert Boyle	English	Hunter, Michael; Centre for Editing Lives and Letters

Detailed information on each corpus

Title	• The Workdiaries of Robert Boyle
Author	Hunter, Michael; Centre for Editing Lives and Letters
Availability	Available for non-commercial use on condition that this header is included in its entirety with any copy distributed. Registration or request via our order form is required. As this resource is restricted in some way, you will have to apply for approval to get a copy.
Languages	English; French; Latin;
Editorial Practice	• Encoding format: TEI XML
OTA keywords	Notebooks
LC keywords	Science -- History -- 17th century -- Archival resources Boyle, Robert, 1627-1691 Boyle, Robert, 1627-1691. Correspondence
Extent	• designation: CollectionText • size: 47 files: ca. 4.66 MB
Creation Date	1998-2001 – initial creation. 2004 – revised edition
Source Description	
Notes	• Title proper taken from AHDS Catalogue Form • The Workdiaries record Robert Boyle's experiments, observations and measurements, information that he was given by travellers and others, and extracts from books. They cover over 40 years of Boyle's scientific work. • Michael Hunter. 2007. The Boyle Papers: Understanding the Manuscripts of Robert Boyle. Aldershot: Ashgate, c2007 • Also available at: http://www.livesandletters.ac.uk/wd/index.html

Related corpora to selected one

Advanced Search Results **1 - 3** of about **3** for **site:ota.oucs.ox.ac.uk/headers/ "OTA keywords" "Notebooks"**. Search took **0.18** seconds.

Sort by date / Sort by relevance

[OTA] The Newton Project [Electronic resource]
... OTA keywords, Biographies. **...** The central focus of the resource is the series of XML encoded transcriptions of Newton's theological works, personal **notebooks**, all draft biographical information about Newton dating from the eighteenth century, and **...**
...ucs.ox.ac.uk/headers/2479.xml - 12k - Cached

[OTA] Research at Oxford on tree-ring dating of oak [Electronic ...
... chronologies developed and single sequences measured; Data is listed as used with cros program, Titles state subject, fletch ref. no., location, years spanned, sapwood years if any. **OTA keywords**, Electronic publications -- Great Britain -- **...**
ota.oucs.ox.ac.uk/headers/0692.xml - 11k - Cached

[OTA] The Workdiaries of Robert Boyle
... Registration or request via our order form is required. As this resource is restricted in some way, you will have to apply for approval to get a copy. Languages, English; French, Latin, Editorial Practice, Encoding format. TEI XML. **OTA keywords ...**
ota.oucs.ox.ac.uk/headers/2529.xml - 11k - Cached

The Oxford Text Archive

- Pros
  - Easy to compare listed corpora because its list also has the additional information on corpora
  - Easy to find corpora belonging to the same category of a selected corpus, which are likely to meet desired characteristics, through the fields ‘OTA keywords’, ‘LC keywords’
- Cons
  - Hard to find a corpus such that some criteria are met

The LDC of the Univ. of Penn.

- http://www.ldc.upenn.edu

- It creates, collects and distributes speech and t ext databases, lexicons, and other resources for research and development purposes

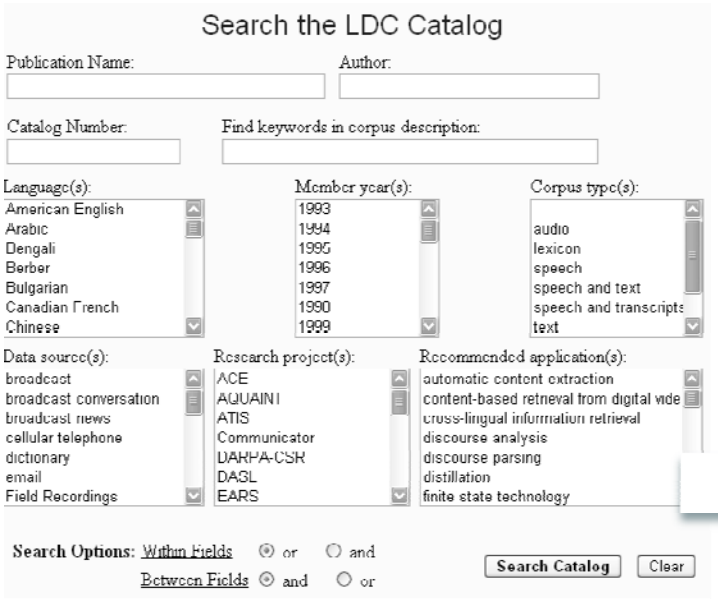
- Features
  - Providing sorted lists and search functionality by several features such as their popularity and recommended applications

The LDC of the Univ. of Penn.

The ten most-purchased corpora

948	LDC93S1	TIMIT Acoustic Phonetic Continuous Speech Corpus
711	LDC96L14	CELEX2
629	LDC2006T13	Web 1T 5-gram Version 1
422	LDC93S10	TIDIGITS
380	LDC94T5	ECI Multilingual Text
307	LDC93S2	NTIMIT
303	LDC99T42	Treebank-3
301	LDC93T3A	TIPSTER Complete
260	LDC94S16	YOHO Speaker Verification
250	LDC2001T02	Message Understanding Conference (MUC) 7

Search functionality



Detailed Information on each corpus

Treebank-3

Item Name: Treebank-3
Authors: Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz and Ann Taylor
LDC Catalog No.: LDC99T42
ISBN: 1-58563-163-9
Data Type: text
Data Source(s): varied
Project(s): GALE, TIDES
Application(s): natural language processing, parsing, tagging
Language(s): English
Language ID(s): eng
Distribution: 1 CD, Web Download
Membership Year(s): 1999
Non-member Fee: US$3150.00
Reduced-License Fee: US$1575.00
Extra-Copy Fee: US$150.00
Non-member License: yes
Online documentation: yes
Licensing Instructions: Subscription Members, Standard Members, Non-Members
Citation: Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz and Ann Taylor
1999
Treebank-3
Linguistic Data Consortium, Philadelphia

The LDC of the Univ. of Penn.

- Pros
  - Easy to find relevant corpora to a specific application through the field 'application(s)'
  - Providing corpora in many languages and from many data sources
- Cons
  - Hard to compare corpora with several factors because the lists has only their descriptions

The European Language Resources Association

- http://www.elra.info

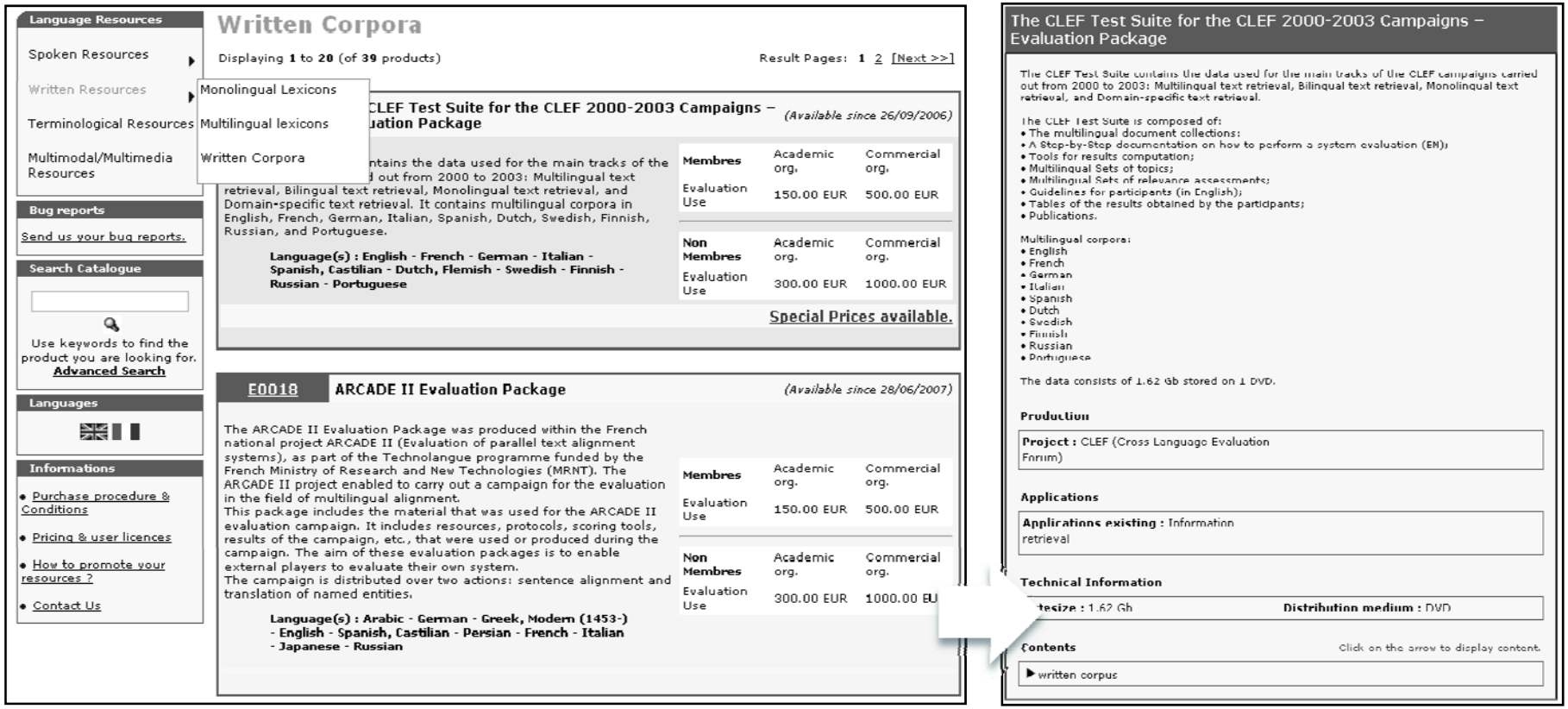
- Missions
  - Make available the language resources for language engineering
  - Evaluate language engineering technologies

- Features
  - Browsing through classification hierarchy
  - Validation processing specialized for the type of corpora
  - Providing specialized sub-lists such as R&D catalogue

The European Language Resources Association

Browsing through the structured list

Detailed information



The European Language Resources Association

- Pros
  - Systemic validation process increasing confidence on the quality of the corpora
  - Specialized sub-lists helping to find corpora with desired characteristics
- Cons
  - Hard to find what kind of corpora is in the list because of its little classification hierarchy