The Extraction Corpus and the Reference Corpus

The introduction

Fundamental aspects concerning the documentation containing the desired information.

Compilation of a special purpose corpus.
The objectives

Upon completion of this lesson, students will be able to:

1. Identify and collect suitable texts for inclusion in a corpus.

2. Apply quality criteria regarding the building of the ad hoc extraction corpus to get a corpus representative of the studied area.

3. Establish a preference order on the reference materials to be consulted.

The outline: 3 main points

1. Documentation containing desired information
2. Compiling a special purpose corpus: electronic resources.
3. Selecting appropriate texts.
The Extraction Corpus and the Reference Corpus

The content - Identifying and Evaluating Specialized Documentation in a Field

When doing terminology research, terminologists need to find information about the available documentation on a given topic in two directions:

- Texts dealing with the subject-field. To acquire knowledge about the field.
- Dictionaries and databases to be consulted. To solve punctual questions about the terms in a later stage.
When doing terminology research, terminologists need to find information about the available documentation on a given topic in two directions:

- **EXTRACTION CORPUS**: Terms that are not known about
- **REFERENCE CORPUS**: Terms that are known about

**The primary function of terminology work is:**
- (a) the transfer of **specialized knowledge**.
- (b) the authentication of related **terminological usage**.

**Essential requirements are:**
- The **acquisition and structuring** of such knowledge by finding the concepts involved.
- The **identification of the terms** that convey this specialized knowledge.

- **the designations of the concepts to be defined**
- **their interrelationships are studied and represented**
To conduct any terminology research purported to reflect the current state of the art:

- keep track of knowledge in a given sphere of activity
- stay abreast of new developments and their impact on communication

This will help you:
- identify basic terminology
- recognize the most recent terminology
- read carefully specialized documentation
- build a network of specialized consultants
- keep informed of relevant topics (simposia, conferences, exhibits)

Documentation can be categorized as follows:

- encyclopaedias
- monographs and technical and academic manuals
- proceedings of congresses and symposia
- specialized and popularized periodicals
- brochures and publicity flyers
- dictionaries, vocabularies, and documentary, terminology, and linguistic databases

Internet sites: best content providers in the area of specialization
The Internet offers information and data all over the world

BUT

- There is so much information.
- Information can appear to be fairly “anonymous”.

SO

- It is necessary to develop skills to evaluate what you find.

Some types of documentation are traditionally preferred over others:

- Original-language documents are preferable to translations.
- Encyclopaedias and other recognized academic documents or works are preferable to brochures and promotional material.
The usefulness of documents is evaluated against criteria such as the following:

- Original language of the documentation
- The text has to belong to the domain
- Glossary
- The structure of the contents
- The publication date
- Specialized - Official - ETC
- Text Nature
- Grammar - VOC-STYLE
- Linguistic quality
- The presence of an up-to-date bibliography
- Expert
- The author’s credentials

The main points so far...

1. We need to collect subject field texts, as well as reference material that will be consulted in a later stage.
2. We need to keep informed from several sources (specialized documentation, experts, symposia, etc.)
3. We can access to several types of documentation (monographs, proceedings, encyclopedias…) and some types are preferred over others.
4. We need to apply several parameters that will allow us to evaluate the usefulness of documents.
### The Extraction Corpus and the Reference Corpus

**The basics about designing a special purpose corpus**: Compiling a special purpose corpus: electronic resources

- **1st task**: map out the design of your ideal corpus
- **2nd task**: identify and collect suitable texts for this corpus

**Practical problems** to build an ideal corpus:
- Not being able to find all the texts you need in electronic form.
- Finding the process of identifying and downloading texts from the Web more time consuming than expected, or
- Not having copyright permission to hold certain texts in your corpus.

These problems mean:

- **Make some adjustments to your ideal design.**
- **Be realistic**
- **Balance:**

For our purposes (final assignment), it is probably not a good use of your time to spend a month constructing the corpus.
The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

❖ Remember...
  ▪ A corpus can still be a useful resource, even if it does not perfectly resemble the ideal corpus that you planned during the design stage.
  ▪ Be aware of any shortcomings that your corpus may have (e.g. some of the texts are a little bit old; some of the authors' credentials are unclear).

Finding information on the Internet is not difficult

❖ BUT

❖ Finding the specific texts we are searching for and of the up most quality is like:

❖ To extract the needle from the haystack, the translator should use a search engine, as it is:

Finding information on the Internet is not difficult, but finding the specific texts we are searching for and of the utmost quality is like: to extract the needle from the haystack, the translator should use a search engine, as it is.
Lesson 3: The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

❖ What does this imply?
   (1) Knowing the functionalities of the search engine we are going to employ.
   (2) Applying the query language that the search engine understands.
   (3) Planning several search strategies.

we will have more possibilities to obtain the desired result

Lesson 3: The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

❖ The Internet offers information and data all over the world
❖ BUT
   ▪ There is so much information.
   ▪ Information can appear to be fairly “anonymous”.
❖ SO
   ▪ It is necessary to develop skills to evaluate what you find.
Lesson 3: The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

- Principal aim of the whole process of information localization and retrieval is:
  - Obtain those documents that best meet the information need
  - To recover information lacking in both documents recovered by the IRS but not relevant to the information need

- What types of search engines do you know?
The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

- **Local:** It searches the information within its Web site, e.g.:

![Image of a website search interface]

---

The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

![Diagram showing 4 types of search engines: Global search engines, active, passive, meta search, domain specific]
The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

**DEFINITION**

Search engine that collects Web pages information by itself. It collects index terms from:
- the text found in pages,
- titles
- HTML meta categories such as "Description" and "Keywords".

**EXAMPLES**

<table>
<thead>
<tr>
<th>Engine</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask Jeeves</td>
<td><a href="http://es.ask.com/?o=312#subject:ask">http://es.ask.com/?o=312#subject:ask</a></td>
</tr>
<tr>
<td>Altavista</td>
<td><a href="http://www.altavista.com">http://www.altavista.com</a></td>
</tr>
<tr>
<td>Google</td>
<td><a href="http://www.google.com">http://www.google.com</a></td>
</tr>
<tr>
<td>Hotbot</td>
<td><a href="http://www.hotbot.com">http://www.hotbot.com</a></td>
</tr>
<tr>
<td>Lycos</td>
<td><a href="http://www.lycos.com">http://www.lycos.com</a></td>
</tr>
<tr>
<td>Netscape Search</td>
<td><a href="http://channels.netscape.com/search/default.isp">http://channels.netscape.com/search/default.isp</a></td>
</tr>
<tr>
<td>Search Engine Colossus</td>
<td><a href="http://www.searchenginecolossus.com/">http://www.searchenginecolossus.com/</a> (directorio de motores y buscadores)</td>
</tr>
</tbody>
</table>
The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

WORKING METHOD

It uses a robot that travels around the Internet, locates Web pages and adds entries to the catalog.

ADVANTAGES/DISADVANTAGES

ADVANTAGES:
* have large catalogs
* are updated frequently (without human intervention).

DISADVANTAGE:
* there are often too many hits, which are not very well organized.
A search engine that allows people to register their Web pages.

**DEFINITION**

**EXAMPLES**

<table>
<thead>
<tr>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://bubl.ac.uk/link/">http://bubl.ac.uk/link/</a></td>
<td>BUBL Link</td>
</tr>
<tr>
<td><a href="http://www.hw.ac.uk/tb/www/archives/pinakes/pinakes.html">http://www.hw.ac.uk/tb/www/archives/pinakes/pinakes.html</a></td>
<td>PINAKES</td>
</tr>
<tr>
<td><a href="http://in0.ac.uk">http://in0.ac.uk</a></td>
<td>RDN</td>
</tr>
<tr>
<td><a href="http://science.gov">http://science.gov</a></td>
<td>Science.gov</td>
</tr>
<tr>
<td><a href="http://scout.cs.wisc.edu/archives">http://scout.cs.wisc.edu/archives</a></td>
<td>Scout Report Archives</td>
</tr>
<tr>
<td><a href="http://www.vlib.org/">http://www.vlib.org/</a></td>
<td>WWW Virtual Library</td>
</tr>
</tbody>
</table>
The Extraction Corpus and the Reference Corpus

Working Method

Once a page is registered with the search engine, the page can be found by queries.

Advantages/Disadvantages

Advantage:
- They tend to be very organized.

Disadvantage:
- Their catalogs are smaller than the active search engines and updating is not automatic, as it needs human intervention.
The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

Metasearch Engine

**DEFINITION**

search engine which uses several search engines simultaneously.

---

The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

**EXAMPLES**

<table>
<thead>
<tr>
<th>Metasearch Engine</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copernic</td>
<td><a href="http://www.copernic.com">http://www.copernic.com</a></td>
</tr>
<tr>
<td>Metacrawler</td>
<td><a href="http://www.metacrawler.com">http://www.metacrawler.com</a></td>
</tr>
<tr>
<td>SurfWax</td>
<td><a href="http://www.surfwax.com">http://www.surfwax.com</a></td>
</tr>
<tr>
<td>Vivisimo</td>
<td><a href="http://vivisimo.com">http://vivisimo.com</a></td>
</tr>
</tbody>
</table>
The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

WORKING METHOD

It sends user requests to several other search engines and/or databases and returns the results from each one

ADVANTAGES/DISADVANTAGES

ADVANTAGE:
* they save effort by searching multiple search engines.

DISADVANTAGES:
* the search can be slow
* they may summarize the data in their own way
* they may present only partial results from each search engine, possibly hiding relevant information
The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

- They can be active or passive search engines or a combination of both.
- They focus their search on major disciplines or specific fields:

<table>
<thead>
<tr>
<th>Scirus</th>
<th>science</th>
<th><a href="http://www.scirus.com">http://www.scirus.com</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ojose</td>
<td>Online Journals Search Engine</td>
<td><a href="http://www.ojose.com/">http://www.ojose.com/</a></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>List of academic databases and search engines</td>
<td><a href="http://goo.gl/Vy3ep">http://goo.gl/Vy3ep</a></td>
</tr>
</tbody>
</table>

Imagine that you have this information need:

I want to obtain a text (not an image) in .pdf or .doc about “checking accounts” or “current accounts”.
This term has to appear in the title or in the body of the recovered Web pages.
I also wish that the word “saving account” appears in the text.
The texts should be posted on any institution/university website.

What would you type in the search box?
The Extraction Corpus and the Reference Corpus

The content – 2. Compiling a special purpose corpus: electronic resources

**Basics**

### operators

- **NO SPACES** between the operator and a word. Do put a space between each operator/word combination.

- **+** asterisk (not supported in Google)
  
  Type an asterisk at the right-hand side of a word to retrieve all the words that start with the one you used. Example: gloss* (for glossario, glossary, glossaire).

- **""** double quotes around a phrase
  
  The quotes create a phrase which must be retrieved exactly as you typed it. Words next to each other and in that order. Example: "solar energy"

- **+** plus sign
  
  A plus sign requires that the word be found in all of the search results.

- **-** minus sign
  
  A minus sign eliminates or excludes any results with that word.

---

**Example:**

<table>
<thead>
<tr>
<th>Search topics</th>
<th>Search statements using operators: quotes, asterisks, plus and minus signs</th>
</tr>
</thead>
<tbody>
<tr>
<td>glossary about transgenic food</td>
<td>glossary +&quot;transgenic food&quot;</td>
</tr>
<tr>
<td>computer assisted translation (not automatic translation)</td>
<td>+&quot;computer assisted translation&quot; +&quot;automatic translation&quot;</td>
</tr>
<tr>
<td>glossary (in English or Spanish) containing the term 'headache'</td>
<td>glos* + 'headache'</td>
</tr>
</tbody>
</table>
The content – 2. Compiling a special purpose corpus: electronic resources

### Boolean

<table>
<thead>
<tr>
<th>Boolean term</th>
<th>Always type the Boolean term in capital letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND</td>
<td>AND requires that the word be found in all of the search results. Example: dictionary AND headaches</td>
</tr>
<tr>
<td>OR</td>
<td>OR is used to broaden your search with alternatives and synonyms. Example: glossary OR vocabulary OR dictionary AND headaches</td>
</tr>
<tr>
<td>NOT</td>
<td>NOT eliminates or excludes any results with that word. You may type AND NOT for the command to work at excluding what you don't want. Example: labour AND NOT childbirth (in Google: labour -childbirth)</td>
</tr>
<tr>
<td>NEAR (not supported in Google)</td>
<td>NEAR usually requires that the words be found within 7-10 words of each other. Example: To look for links to pages about mobile phones and ISO standards, you could try [mobile phones] near:ISO</td>
</tr>
</tbody>
</table>

### Field-Limiting

**Field-limiting operator for Google**

- **related**: search pages similar to the typed page
  - Example: the-universidad-bilbao
- **allintitle**: all the keywords are in the title.
  - Example: allintitle: the-universidad-bilbao
- **inurl**: search in URL. E.g.: inurl:chelo vargas
- **allinurl**: all keywords are in URL
  - Example: allinurl: the-universidad-bilbao
- **link**: pages related to the typed page. E.g. link: http://www.lulacs.es
- **site**: search in a site
  - Example: <reglamento site:ua.es> (it will search the keyword in the University of Alicante webpage in Spain). You can also search in a domain, such as: .iai, .com, .edu, .org, etc.
- **filetype**: search for a document in a specific format, such as.pdf, .doc, .ppt, etc.
  - Example: headaches filetype.pdf
The content – Compiling a special purpose corpus: electronic resources

**RESULTING SEARCH QUERY:**

```
“saving account” “checking account” OR “current account” filetype:pdf OR filetype:doc site:edu OR site:org
```

The main points so far...

1. Search engines can be global and local.
2. Global engines can be divided into: active, passive, metasearch, and domain-specific.
3. There are 3 types of operators that can be combined with one or more keywords: basic, boolean, and field-limiting.