



**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 Extraction Corpus and the Reference Corpus

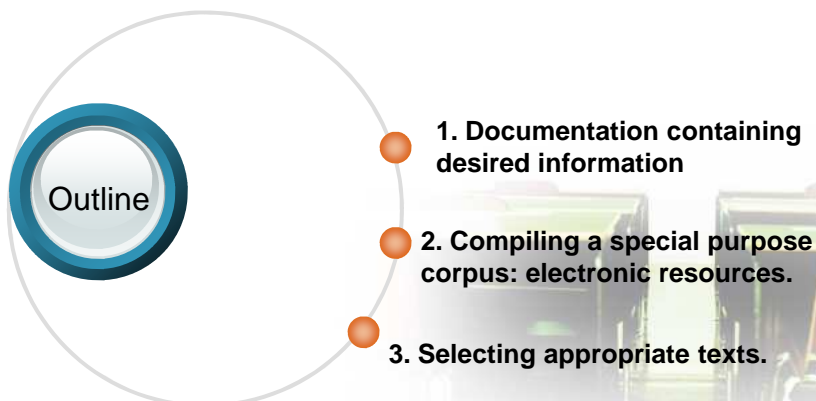
### The objectives

❖ **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 Extraction Corpus and the Reference Corpus

### The outline: 3 main points



## The Extraction Corpus and the Reference Corpus

### The content - Identifying and Evaluating Specialized Documentation in a Field



DOCTOR, I LOOKED UP  
MY SYMPTOMS  
ON THE INTERNET  
AND I THINK I MIGHT BE DEAD!

DON'T BELIEVE EVERYTHING  
YOU READ ON THE NET

## The Extraction Corpus and the Reference Corpus

### The content

**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.**

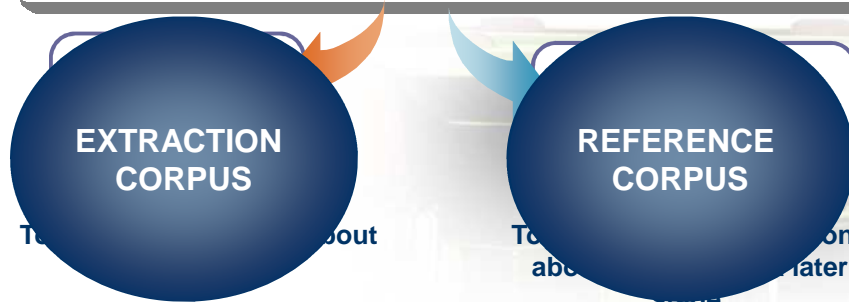
**Dictionaries and databases to be consulted.**

**To solve punctual questions about the terms in a later stage.**

## The Extraction Corpus and the Reference Corpus

### The content

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



## The Extraction Corpus and the Reference Corpus

### The content - The content - Documentation Containing Desired Information

#### ❖ 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.



are

the designations of the concepts to be defined

their interrelationships are studied and represented

## The Extraction Corpus and the Reference Corpus

The content - The content - Documentation Containing Desired Information

❖ **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

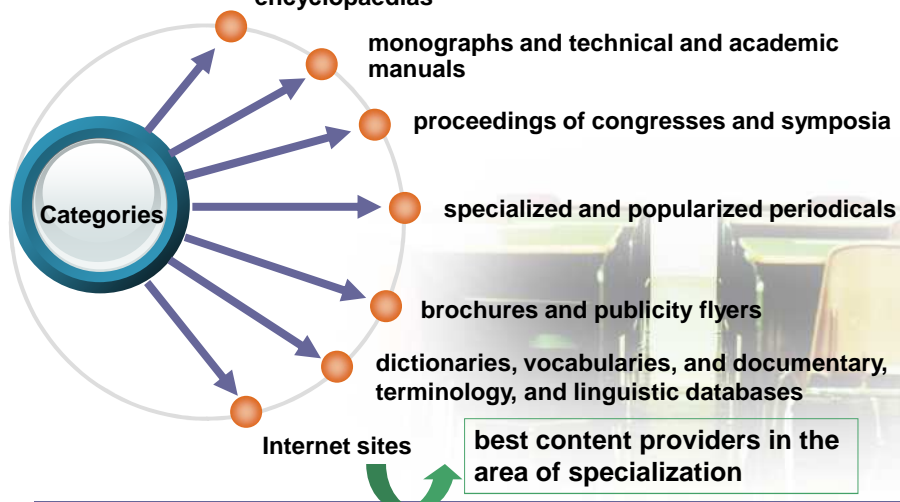


keep informed of relevant topics (symposia, conferences, exhibists)

## The Extraction Corpus and the Reference Corpus

The content - The content - Documentation Containing Desired Information

**Documentation can be categorized as follows:**



## The Extraction Corpus and the Reference Corpus

The content - The content - Documentation Containing Desired Information

❖ **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.



Its selection must always be adapted to your particular research situation

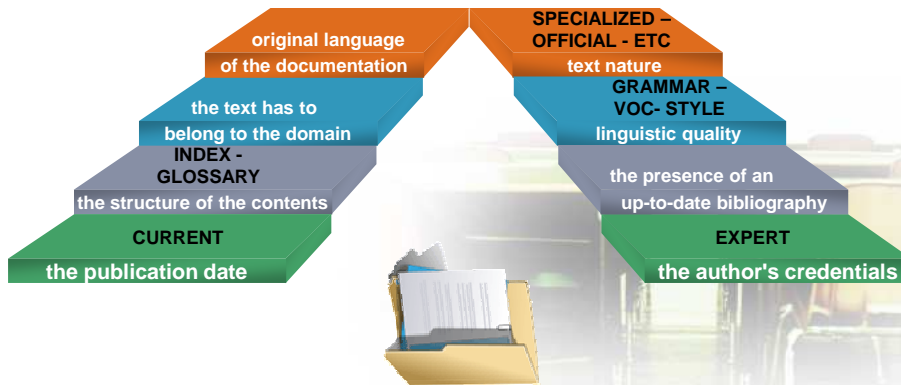
❖ **Some types of documentation are traditionally preferred over others:**

- **Original-language documents** are preferable to translations.
- **encyclopaedias** and other recognized **academic documents** or works are preferable to brochures and promotional material.

## The Extraction Corpus and the Reference Corpus

The content - The content - Documentation Containing Desired Information

The usefulness of documents is evaluated against criteria such as the following



## The Extraction Corpus and the Reference Corpus

The content –Documentation Containing Desired Information

The main points so far...

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.



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.)

## The Extraction Corpus and the Reference Corpus

The basics about designing a special purpose corpus

2. 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



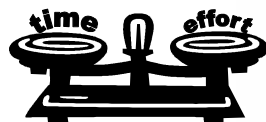
- Not having copyright permission to hold certain texts in your corpus.

## The Extraction Corpus and the Reference Corpus

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

These problems mean:

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



For our puposes (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).



keep these in mind when interpreting the data

## The Extraction Corpus and the Reference Corpus

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



- ❖ **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:**

## The Extraction Corpus and the Reference Corpus

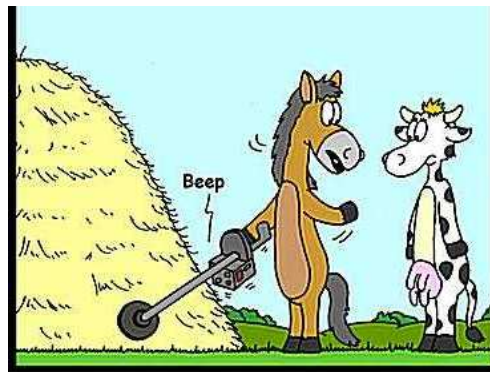
The content – 2. Compiling a special purpose corpus: electronic resources



Chelo Vargas - <http://www.ua.es/personal/chelo.vargas/index.html>

## The Extraction Corpus and the Reference Corpus

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



You were right: There's a needle in this haystack...

Chelo Vargas - <http://www.ua.es/personal/chelo.vargas/index.html>

## 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

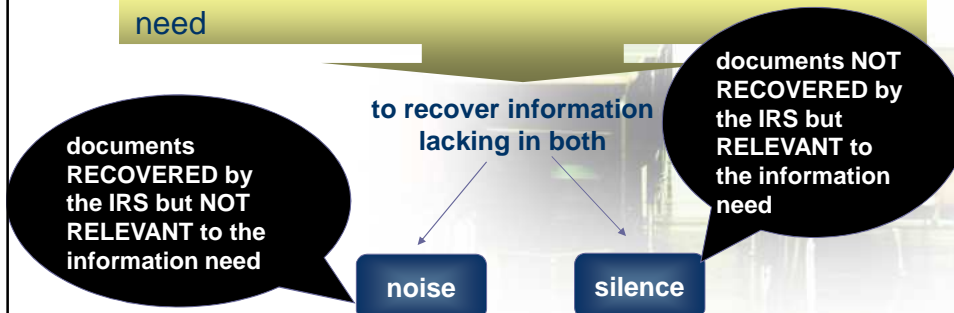
- It is necessary to develop skills to evaluate what you find.

### 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



### The Extraction Corpus and the Reference Corpus

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

❖ **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.:



## The Extraction Corpus and the Reference Corpus

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

Global search engines

4 types of 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

#### Active Search Engine

##### 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".**

## The Extraction Corpus and the Reference Corpus

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

#### Active Search Engine

##### EXAMPLES

Ask Jeeves	<a href="http://es.ask.com/?o=312#subject:ask pg:1">http://es.ask.com/?o=312#subject:ask pg:1</a>
Altavista	<a href="http://www.altavista.com">http://www.altavista.com</a>
Google	<a href="http://www.google.com">http://www.google.com</a>
Hotbot	<a href="http://www.hotbot.com">http://www.hotbot.com</a>
Lycos	<a href="http://www.lycos.com">http://www.lycos.com</a>
NetScape Search	<a href="http://channels.netscape.com/search/default.jsp">http://channels.netscape.com/search/default.jsp</a>
Search Engine Colossus	<a href="http://www.searchenginecolossus.com/">http://www.searchenginecolossus.com/</a> (directorio de motores y buscadores)

## The Extraction Corpus and the Reference Corpus

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

Active Search Engine

### WORKING METHOD

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

## The Extraction Corpus and the Reference Corpus

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

Active Search Engine

### 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.

## The Extraction Corpus and the Reference Corpus

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

#### Passive Search Engine

##### DEFINITION

A search engine that allows people to register their Web pages.

## The Extraction Corpus and the Reference Corpus

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

#### Passive Search Engine

##### EXAMPLES

BUBL Link	<a href="http://bubl.ac.uk/link/">http://bubl.ac.uk/link/</a>
PINAKES	<a href="http://www.hw.ac.uk/libwww/irn/pinakes/pinakes.html">http://www.hw.ac.uk/libwww/irn/pinakes/pinakes.html</a>
RDN	<a href="http://rdn.ac.uk">http://rdn.ac.uk</a>
Science.gov	<a href="http://science.gov">http://science.gov</a>
Scout Report Archives	<a href="http://scout.cs.wisc.edu/archives">http://scout.cs.wisc.edu/archives</a>
WWW Virtual Library	<a href="http://www.vlib.org/">http://www.vlib.org/</a>



## The Extraction Corpus and the Reference Corpus

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

Passive Search Engine

### WORKING METHOD

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



## The Extraction Corpus and the Reference Corpus

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

Passive Search Engine

### 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

### Metasearch Engine

#### EXAMPLES

Copernic	<a href="http://www.copernic.com">http://www.copernic.com</a>
Metacrawler	<a href="http://www.metacrawler.com">http://www.metacrawler.com</a>
SurfWax	<a href="http://www.surfwax.com">http://www.surfwax.com</a>
Vivisimo	<a href="http://vivisimo.com/">http://vivisimo.com/</a>

## The Extraction Corpus and the Reference Corpus

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

### Metasearch Engine

#### WORKING METHOD

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



## The Extraction Corpus and the Reference Corpus

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

### Metasearch Engine

#### ADVANTAGES/DISADVANTAGES

##### ADVANTAGE:

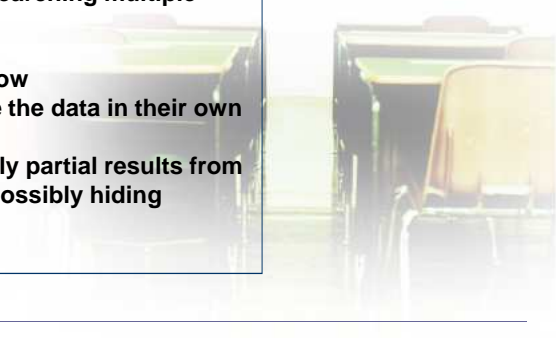
\* they save effort by searching multiple search engines.

##### 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

### Domain-specific

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

Scirus	science	<a href="http://www.scirus.com">http://www.scirus.com</a>
Ojose	Online JOURNALS Search Engine	<a href="http://www.ojose.com/">http://www.ojose.com/</a>
Wikipedia	List of academic databases and search engines	<a href="http://goo.gl/Vy3ep">http://goo.gl/Vy3ep</a>

## The Extraction Corpus and the Reference Corpus

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

### ❖ 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	<b>NO SPACES between the operator and a word. Do put a space between each operator/word combination.</b>
* 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: glos* (for glosario, glossary, glossaire, glosa)
"....." 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.
- a minus sign	a minus sign eliminates or excludes any results with that word

## The Extraction Corpus and the Reference Corpus

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

**Basics**

**Example:**

Search topics	Search statements using operators: quotes, asterisks, plus and minus signs
glossary about transgenic food	+glossary +"transgenic food"
computer assisted translation (not automatic translation)	+"computer assisted translation" -"automatic translation"
glossary (in English or Spanish) containing the term 'headache'	+glos* +headache

## The Extraction Corpus and the Reference Corpus

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

Boolean terms	Always type the Boolean term in capital letters
<b>AND</b>	<b>AND</b> requires that the word be found in all of the search results Example: dictionary <b>AND</b> headaches
<b>OR</b>	<b>OR</b> is used to broaden your search with alternatives and synonyms. Example: glossary <b>OR</b> vocabulary <b>OR</b> dictionary <b>AND</b> headaches
<b>NOT</b>	<b>NOT</b> Eliminates or excludes any results with that word You must type <b>AND NOT</b> for the command to work at excluding what you don't want Example: labour <b>AND NOT</b> childbirth (in Google: labour -childbirth)
<b>NEAR</b> (not supported in Google)	<b>NEAR</b> usually requires that the words be found within 7-10 words of each other Example: To look for links to pages about mobile booths and ISO standards, you could try [ <a href="#">mobile booth</a> * <b>NEAR</b> <a href="#">iso</a> ]

## The Extraction Corpus and the Reference Corpus

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

**Field-limiting operator for Google**

**related:** search pages similar to the typed page  
**intitle:** search in webpage title.  
 Example: intitle:universidad alicante  
**allintitle:** all the keywords are in the title.  
**inurl:** search in URL. E.g.: inurl:chelo vargas  
**allinurl:** all keywords are in URL  
**link:** pages related to the typed page. E.g.  
 link:http://www.iulma.es  
**site:** search in a site  
 For 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: .es; .com; .edu; .org, etc.)  
**filetype:** search for a document in a specific format, such as.pdf; .doc; .ppt, etc.  
 Ex: <headache filetype:pdf>

## The Extraction Corpus and the Reference Corpus

The content – 2. 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 Extraction Corpus and the Reference Corpus

The content – Compiling a special purpose corpus: electronic resources

### The main points so far...

3. There are 3 types of operators that can be combined with one or more keywords: basic, boolean, and field-limiting.



1. Search engines can be global and local.

2. Global engines can be divided into: active, passive, metasearch, and domain-specific.