

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The introduction

- Close relationship between terminology and computer science and its influence on methodology.
- Terminology processing is corpus-based: parameters for designing a special-purpose corpus.
- Technologies for terminology-related tasks: term extractors and terminology management systems.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The outline

1. Terminology and Computing: an overview
2. Electronic and Special-purpose corpora: parameters for their design and representativeness
3. Systems for term detection and extraction.
4. Terminology Management Systems in Translation Memories

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – What is terminology?

❖ Terminology

science

The scientific study of the concepts and terms found in special languages
(ISO 1087: 2000, 12)



practice

The guidelines used in terminographic work, referring to its methodology



object

The set of terms of a particular special subject



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

❖ **The effect of computer science on the methodology of terminology:**

- in the use of previously recorded **electronic corpora**;
- the exploitation of terminological, and knowledge **databases**.



❖ **The effect of computer science on the methodology of terminology:**

- **RESULT:**
 - Terminology and lexicography have come closer together
 - Change in the way terminological methodology approaches texts
 - Increasingly important role to the linguistic context of terms
 - The ability to access large databases has therefore brought about a change in the focus of terminology work.

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content - Computer science and terminological methodology

❖ The effect of computer science on the methodology of terminology:

▪ RESULT:

- Terminology and lexicography have come closer together

Terminology and lexicography are in theory two quite different disciplines, yet in practice they are headed towards a single methodology because of their common use of computer applications. (Sager, 1990)

- The ability to access large databases has therefore brought about a change in the focus of terminology work.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content - Computer science and terminological practice

❖ Large amount of information computers have made available to:

- Translators
- Terminologists

can now be much more confident about the decisions they make about terms.

❖ The ability to access huge amounts of information stored in distant databanks also makes terminology work:

- Much easier
- More complex

Terminology has gone from being an art to being a technique (Sager, 1990)

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

The content - basic points in terminography

❖ 5 basic points where computers can play highly significant role for terminologists (Cabrè, 1998):

1. Selecting documentation, prior to beginning work
2. creating the corpus and isolating and extracting data
3. writing the entry
4. checking the information in the entry
5. ordering the terminological entries

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

The content - selecting documentation

1. Selecting documentation

Terminologists now have access to various types of databanks with general, background information;

1.1 document banks (publications and dictionaries about the subject)

1.2 text banks of corpora of technical texts;

1.3 terminological banks with lists of terms about the subject and linguistic and interlinguistic information about each term.

2. Creating the corpus and isolating and extracting data

3. Writing the entry

4. Checking the information in the entry

5. Ordering the terminological entries.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content - creation of the corpus

1. Selecting documentation

2. Creating the corpus and isolating and extracting data

2.1 Terminologists can **select electronic texts** and incorporate them into a text databank

2.2 Paper texts can be put into electronic form with the aid of an **optical scanner and an OCR**

2.3 Electronic **texts** can be **analyzed** by a semi- or fully automated extraction programs to **isolate and extract terms**

3. Writing the entry

4. Checking the information in the entry

5. Ordering the terminological entries.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – writing the entry

1. Selecting documentation

2. Creating the corpus and isolating and extracting data

3. Writing the entry

3.1 For drafting term records: terminologist use computer files for writing terminological records by transferring some of the information from the reference text files (e.g. entry, source, context, definition, etc.)

3.2 Once the record is complete, it can be **edited** in its entirety or just partially, or it can be **combined** with records from other sources.

3.3 Terminologists can also **control** information about **cross-references** and **equivalences**.

4. Checking the information in the entry

5. Ordering the terminological entries.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – checking the information

1. Selecting documentation
2. Creating the corpus and isolating and extracting data
3. Writing the entry
- 4. Checking the information in the entry**

For checking and finalizing the entries, terminologists **go back to databases** and **transfer** any missing **information** to the record.

5. Ordering the terminological entries.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content - publishing

1. Selecting documentation
2. Creating the corpus and isolating and extracting data
3. Writing the entry
4. Checking the information in the entry
- 5. Ordering the terminological entries**

For publishing, computers allow terminologists to present the information in various formats (paper, disk, tape, file, optical disk), to include the data that are best suited to each case.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content - Electronic and special-purpose corpora



❖ A search for a single term not listed in a dictionary involved:

- Selecting bibliography
- Consulting experts in the field



❖ Access terminological databases and special text databases has markedly changed working methods.



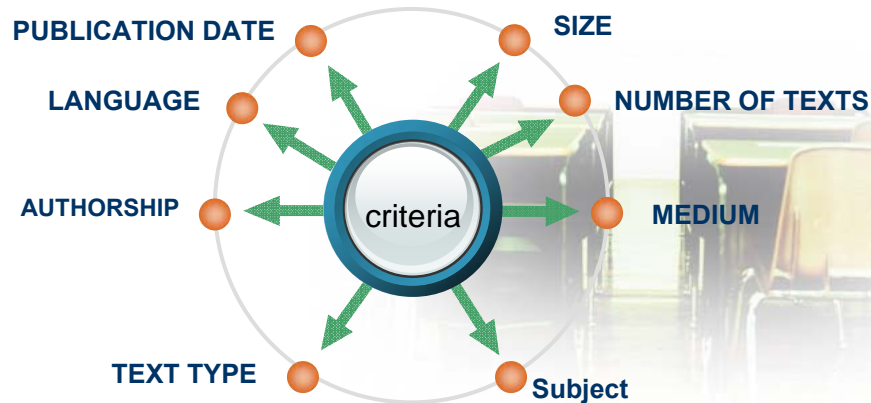
❖ Today, systematic terminology is based on researching a representative corpus of texts.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content - Electronic and special-purpose corpora

❖ Basic guidelines to help in the design of special-purpose corpora:



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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content - Electronic and special-purpose corpora

❖ Corpus:

- collections of texts that have been put together according to specific criteria

❖ These criteria are determined by:

- Users' needs
- The goals of the project

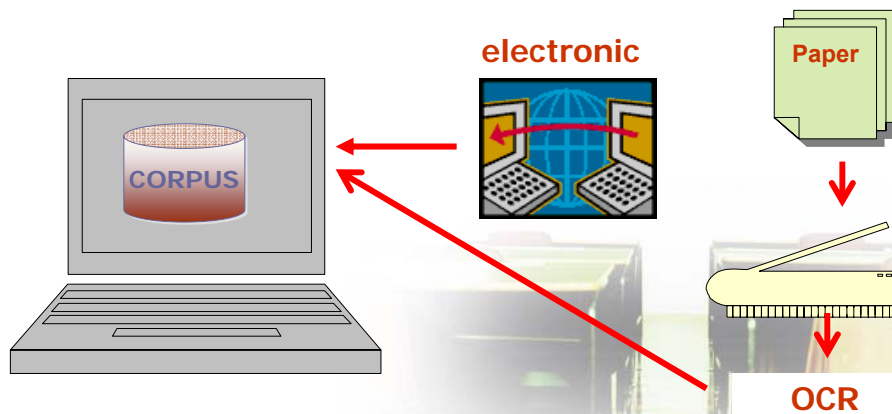
a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language (Sinclair, 1996)

(adopted by the European Advisory Group on Language Engineering Standard)

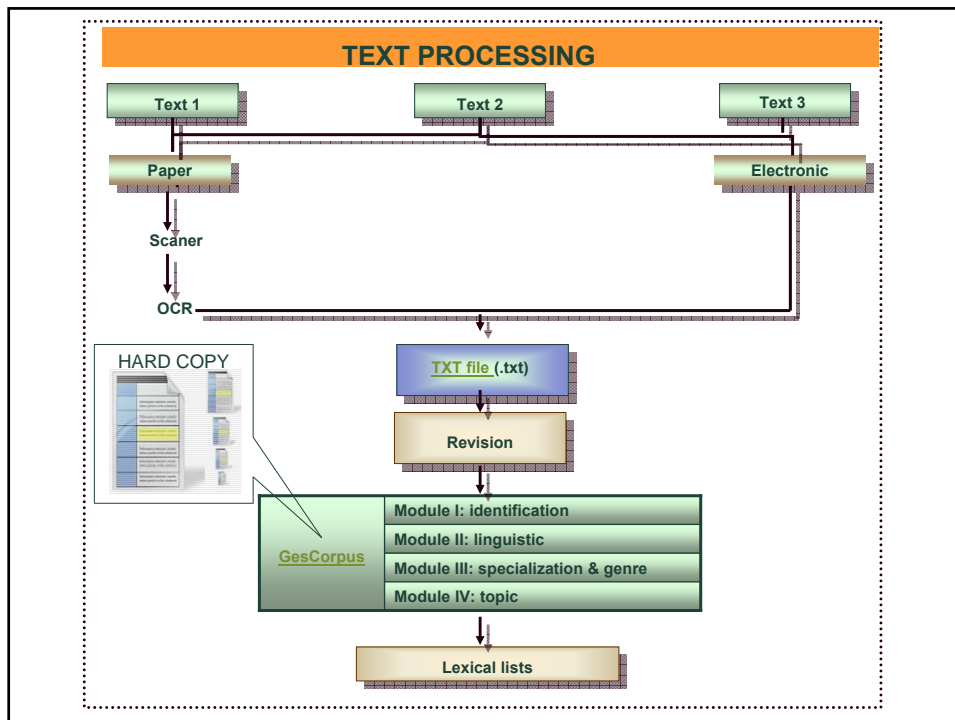
Chelo Vargas - <http://www.ua.es/personal/chelo.vargas/index.html>

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – CORPUS



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



Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

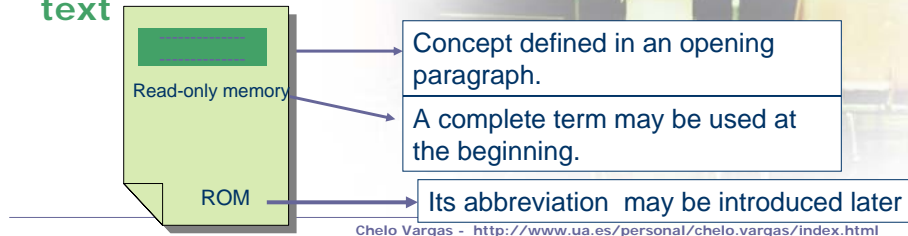
The content – CORPUS SIZE

- ❖ Corpus is often referred to as a 'large' collection of text (but 'large' is rather vague)
- ❖ There are no hard and fast rules that can be followed to determine the ideal size of a corpus.
- ❖ Instead, the decisions are based on factors such as:
 - The needs of the project
 - The availability of data
 - The amount of time we have
- ❖ It is important not to assume that bigger is always better
- ❖ Useful information can be found in a small but well-designed corpus customized to meet the user's needs

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

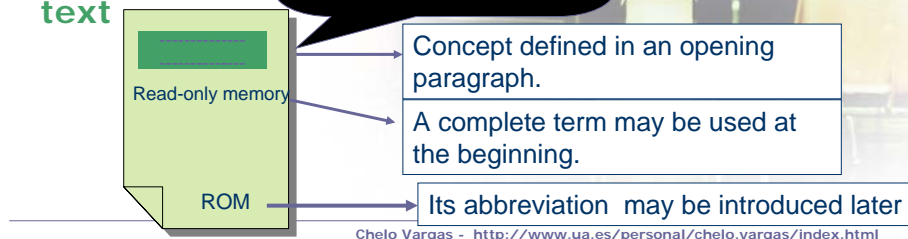
The content – TEXT EXTRACTS vs. FULL TEXTS

- ❖ In some LSP corpora, the size of any given text sample are limited
 - E.g. Lancaster-Oslo-Bergen corpus: extracts for only 2.000 words.
- ❖ In terminology, the concepts, terms, patterns and contexts may appear in any section of text



The content – TEXT EXTRACTS vs. FULL TEXTS

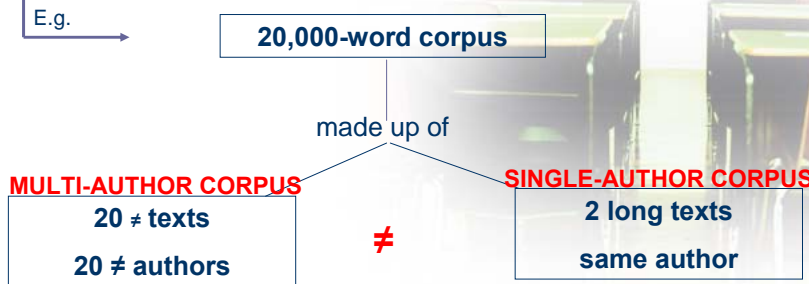
- ❖ In some LSP corpora, the size of any given text sample are limited
 - E.g. Lancaster-Oslo-Bergen corpus: extracts for only 2.000 words.
- ❖ In terminology, the concepts, terms, patterns and contexts may appear in any section of text



The content – NUMBER OF TEXTS

❖ It is also important to consider:

- How many different texts are to be included in the corpus.
- How many of these texts have been written by different authors.

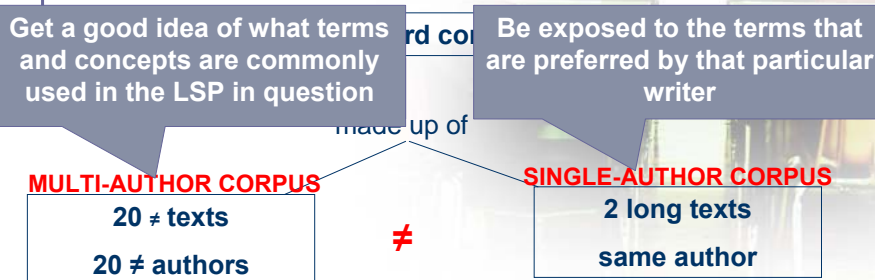


Chelo Vargas - <http://www.ua.es/personal/chelo.vargas/index.html>

The content – NUMBER OF TEXTS

❖ It is also important to consider:

- How many different texts are to be included in the corpus.
- How many of these texts have been written by different authors.



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

The content - MEDIUM

❖ Refers to whether a text was originally prepared as:

- Written texts
- Spoken texts (transcribed)

TRANSLATORS &
TERMINOLOGISTS

The content - SUBJECT

❖ The texts to be included in a corpus need to be about the specialized subject under study

- This is not always as easy as it sounds!

❖ Why?

- Many specialized subjects are multidisciplinary (biochemistry)

❖ Where does biochemistry stop and chemistry begin?

- it can be difficult to know where one subject field ends and the next begins.

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – TEXT TYPE



TRANSLATOR:

- Hired to translate a **research paper for asthma**:
- s/he will benefit from a corpus with **other research papers** on asthma
- These texts will contain examples of the **vocabulary** and **style** that are **appropriate** to this **text type**

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – TEXT TYPE



TRANSLATOR:

- Hired to translate a **research paper for asthma**:
- s/he will benefit from a corpus with **other research papers** on asthma
- These texts will contain examples of the **vocabulary** and **style** that are **appropriate** to this **text type**

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

The content - AUTHORSHIP

❖ To obtain a corpus with authentic LSP material:

- The author of each text should be an acknowledge subject field expert

suitable educational and/or professional background in the discipline about which they write

recognized by their peers as having the level of expertise required to write about the particular subject

The content - AUTHORSHIP

❖ To obtain a corpus with authentic LSP material:

- The author of each text should be an acknowledge subject field expert
- It may be easy to identify the author of a printed document, but when dealing with the Web it may be difficult.
- On the Web:
 - Someone's personal home page may be less reliable than
 - A text that is posted on the web site of a recognized professional organization.

The content - LANGUAGE

- ❖ Using original language material will provide the user with authentic examples of typical LSP use

The content – PUBLICATION DATE

- ❖ For the current state of a field (both at the linguistic and conceptual level)

**INCLUDE MAINLY UP-TO-DATE
TEXTS IN THE CORPUS**

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content –Electronic and special-purpose corpora

The main points so far...

2. A corpus is a collection of (electronic) texts that have been put together according to specific criteria



1. The basic guidelines that help us in the design of a corpus are: size, number of texts, medium, subject, text type, authorship, language, publication date

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content - Systems for term detection and extraction

❖ Term extraction tools (also *term recognition/identification tools*) can be:

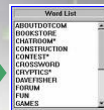
- **Monolingual:** attempt to analyze a text or corpus to identify candidate terms.
- **Bilingual:** analyze existing source texts along with their translations in an attempt to identify potential terms and their equivalents.



program



Initial extraction



Resulting list of candidates

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

The content - Systems for term detection and extraction

❖ **Term extraction tools (also *term recognition/identification tools*) can be:**

- **Monolingual:** attempt to analyze a text or corpus to identify candidate terms.
- **Bilingual:** analyze existing source texts along with their translations in an attempt to identify potential terms and their equivalents.

Term extraction is **COMPUTER-AIDED**
rather than
FULLY AUTOMATIC



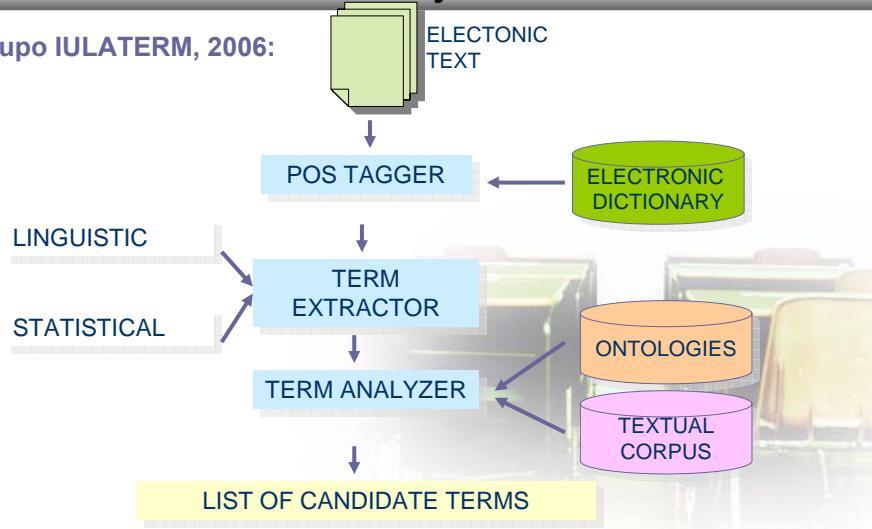
The content: Term extraction

❖ **There are three main approaches to term extraction (Estopà, 1999):**

- **Linguistic:**
 - Attempt to identify word combinations that match certain part-of-speech patterns (e.g., ADJ+N or N+N).
 - heavily language-dependent (term formation patterns differ from language to language).
- **Statistical:**
 - look for repeated sequences of lexical items.
 - Language-independent.
 - The amount of “noise” is relatively high.
- **Hybrid:**
 - Both approaches (linguistic + statistical).

The content: Term extraction - hybrid

Grupo IULATERM, 2006:



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

The content: Term extraction – STATISTICAL

❖ **Concordancers are used**

- in the analysis of **literary texts** to identify:
 - Vocabulary usage peculiarities
 - Turns of phrase of a writer
- When used in **terminology**:
 - Can **count** and **list** the occurrences of a given term in a text

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content: Term extraction – STATISTICAL

The screenshot shows a software window titled "all_en.lst" with a menu bar (File, Edit, View, Compute, Settings, Help) and a toolbar. The main area contains a table with the following data:

N	Word	Freq.	%	Texts	%	Lemmas	Set
1	STONE	4.765	0,89	183	93,37		
2	USED	1.752	0,33	163	83,16		
3	TEST	1.712	0,32	86	43,88		
4	NATURAL	1.490	0,28	148	75,51		
5	SURFACE	1.351	0,25	135	68,88		
6	ROCK	1.149	0,21	92	46,94		
7	WATER	1.113	0,21	121	61,73		
8	MATERIAL	1.023	0,19	142	72,45		
9	MARBLE	1.006	0,19	122	62,24		
10	USE	1.003	0,19	143	72,96		
11	MATERIALS	961	0,18	126	64,29		
12	SPECIMENS	782	0,15	51	26,02		
13	METHODS	749	0,14	97	49,49		
14	ROCKS	747	0,14	77	39,29		
15	STONES	720	0,13	133	67,86		
16	SLABS	690	0,13	87	44,39		
17	GRANITE	671	0,12	92	46,94		
18	SPECIMEN	668	0,12	36	18,37		
19	TYPE	654	0,12	107	54,59		

A blue box with the text "WORDLIST" is overlaid on the right side of the table. At the bottom of the window, there are tabs for "frequency", "alphabetical", "statistics", "filenames", and "notes". The status bar shows "16.789 Type-in".

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

The screenshot shows a software window titled "all_en_lematizada.lst" with a menu bar (File, Edit, View, Compute, Settings, Help) and a toolbar. The main area contains a table with the following data:

N	Word	Freq.	%	Texts	%	Lemmas	Set
16	SLABS	690	0,13	87	44,39		
17	GRANITE	852	0,12	92	46,94	granite[671] granites[181]	
18	SPECIMEN	1.450	0,12	36	18,37	specimen[668] specimens[782]	
19	TYPE	1.021	0,12	107	54,59	type[654] types[367]	
20	DIFFERENT	643	0,12	131	66,84		
21	BUILDING	946	0,12	108	55,10	building[633] buildings[313]	
22	VARIETIES	610	0,11	47	23,98		
23	STRENGTH	614	0,11	84	42,86	strength[602] strengths[12]	
24	SIZE	770	0,11	119	60,71	size[600] sizes[170]	
25	STANDARD	739	0,11	90	45,92	standard[592] standards[147]	
26	RESISTANCE	567	0,11	90	45,92		
27	BLOCKS	564	0,10	79	40,31		
28	AESTHETIC	519	0,10	43	21,94	aesthetic[513] aesthetics[6]	
29	PRODUCTION	522	0,10	85	43,37	production[512] productions[10]	
30	COLOR	615	0,09	43	21,94	color[508] colors[107]	
31	PREN	505	0,09	28	14,29		
32	NUMBER	521	0,09	108	55,10	number[502] numbers[19]	
33	CUT	591	0,09	92	46,94	cut[501] cuts[90]	
34	MADE	492	0,09	130	66,33		

At the bottom of the window, there are tabs for "frequency", "alphabetical", "statistics", "filenames", and "notes". The status bar shows "16.789 Type-in SLABS".

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content: CONCORDANCERS

- ❖ They can also show the term together with the words that precede it and follow

Line	Text	Set	Tag	Word #	t. #	os. #	# os.
542	A condition of stone in which the outer surface of the stone splits apart into			533	23 8%	0	0 8%
543	sufficient, may cause parts of the outer surface of the masonry to spill off or			1.841	72 0%	0	0 9%
544	which forces off the outer surface or layers of masonry. Spalling			1.545	63 7%	0	0 4%
545	of the slot and its depth from the panel surface. Some failures of slots			2.364	96 0%	0	0 3%
546	showing uneven but generally parallel surface patterns. Certain types of gneiss			5.632	242 4%	0	0 3%
547	increases with increasing particle surface area and decreasing size of			224	10 2%	0	0 0%
548	cannot prevent the pavement surface from being coated with ice or a			304	12 6%	0	0 0%
549	Channels may form part of the paving surface, e.g. by creating a dish			4.537	206 2%	0	0 3%
550	correlation with one another. The peak surface temperatures depend on the			3.554	174 1%	0	0 6%
551	is to determine the maximum load per surface unit that a sample is able to bear			998	29 4%	0	0 2%
552	developing a green color and a pitted surface and it also alters in sheltered			68	3 9%	0	0 0%
553	plane: 1) A planar or nearly planar surface that visibly separates the			734	13 9%	0	0 7%
554	the steps needed to form a plane surface from a rough block, as it requires			3.922	183 4%	33	1%
555	pegs. HAND CUTTING A PLANE SURFACE Starting with the rough stone			4.109	193 9%	34	5%
556	supported. Hand-cutting a plane surface Stoneworking, especially the			3.904	183 4%	33	7%
557	Thermal-treatment surface Planed surface Surface treated with chemical			325	12 7%	0	0 5%
558	of the crystals visible on a polished surface. ABSOLUTE BLACK GRANITE			8.732	454 0%	0	0 7%
559	for exteriors, since the polished surface does not last long. Aesthetic			5.619	419 5%	0	0 3%
560	and serpentine will keep a polished surface in exterior work. Polished marble			1.796	84 7%	0	0 7%
561	one is taken advantage of. A polished surface is even considered as a			707	13 4%	0	0 3%

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content: CONCORDANCERS

- ❖ They can also show the term together with the words that precede it and follow

N	L5	L4	L3	L2	L1	Centre	R1	R2
1	THE	THE	OF	OF	THE	ROCK	IS	THE
2	OF	AND	THE	A	OF		IN	IS
3	AND	OF	A	THE	A		AND	OF
4	TO	IN	AND	IN			WITH	A
5	A	A	WHICH	TO	IGNEOUS		MASS	AND
6	IN	IS	TO	AND	SEDIMENTARY		A	TO
7	IS	TO	IN	OR	PLUTONIC		WHICH	IN
8	ROCK	ON	FROM	FOR	VOLCANIC		THE	WHICH
9	OR	ARE	IS	GRAINED	AND		FORMING	ARE
10	AS MECHANICAL	FOR	GIVES	ORNAMENTAL			OR	MINERALS
11	AN	OTHER	RIFT	GIVE	CARBONATE		TYPES	BY
12	2	BE	RESISTANCE	IF	NATURAL		CONSISTING	BE
13	WITH	MINERALS	ROCK	ROCK	THIS		COMPOSED	STONE
14	WHICH	FOR	PROPERTIES	INTO	COMMON		TO	TERMS
15	ON	IGNEOUS	1	WHEN	IN		THAT	IT
16	STRUCTURE	GRAINED	COMPOSITION	AN	OTHER		QUALITY	OR EARA
17	BY	STONE	FINE	FROM	SOLID		OF	MINERAL
18	COARSE	1	PATTERN	ON	PHOSPHATE		FRAGMENTS	CAN

to.vargas/index.html

KWIC: Mother terms

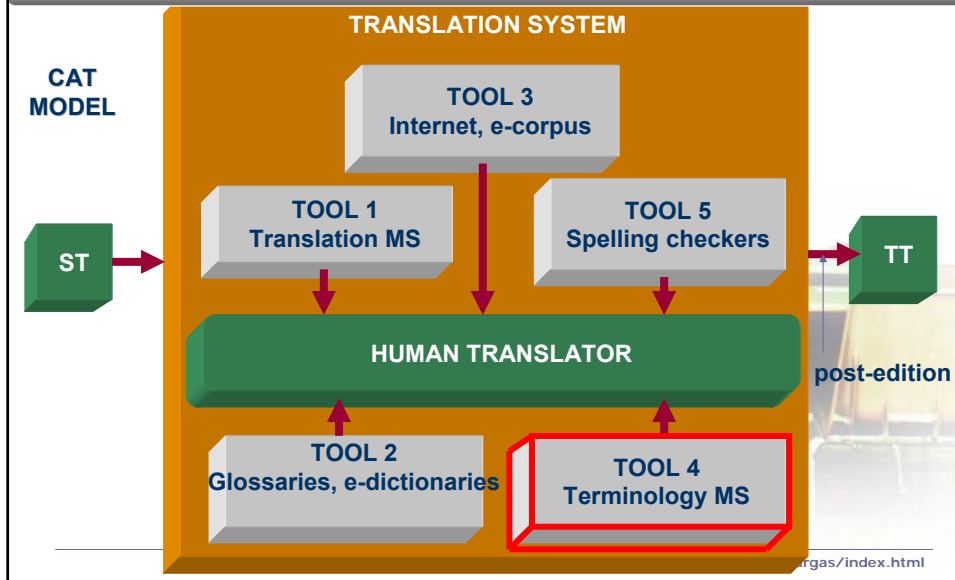
clasto

N	Concordance
1	o con matriz arcillosa en la que se ven algunos clastos. Accesorios: minerales de arcilla, opac
2	nas areniscas de grano fino-medio, con algunos clastos carbonatados y restos fósiles. Su colo
3	sicos. Es frecuente la presencia de zonas con clastos de dolomía que presentan una matriz m
4	e calcita. Mármol de calcita heterogranular con clastos redondeados de cuarzo. Minerales acc
5	arillo Espejón. Fig. 3.-Calizas lacustres con intraclastos y marmorización sinsedimentaria.
6	los clastos que presentan: • Conglomerados, clastos > 2mm. • Samitas, arenas consolida
7	o-medio que presenta una mayor proporción de bioclastos y granos de cuarzo. Son de color a
8	ma rojizo variable. Micrita con lamelibranchios e intraclastos de esparita. Accesorios: minerales
9	rita (Cehegin): roca carbonática con moluscos e intraclastos. Minerales accesorios: arcilla y cua
10	on las de grano grueso, siempre que no existan microclastos ni porfiroblastos, es decir, siemp
11	ras) cementados por calcita espática. Tanto los clastos como el cemento contienen el óxido
12	de cuarzo y feldespatos del 15 %, estando los bioclastos constituidos por foraminíferos, globi
13	cas carbonatadas ignorando el tamaño de los clastos, de los materiales amorfos o de los org
14	anitas se clasifican por la composición de los clastos de mayor tamaño que componen el esq
15	ue conocidas como areniscas, si predominan los clastos entre 2 mm y 0,06 mm. • Lutitas, si l
16	tipo caliche que puede acabar englobando los clastos. II.- Rocas metamórficas: Las rocas m
17	or tamaño que componen el esqueleto, y de los clastos menores que forman la matriz, así com
18	ificaciones, descritas como cataclásticas si los clastos no están alargados y como milonítica
19	calcita microesparítica de grano muy fino, los bioclastos, que suponen entre el 50 y 60% de l
20	sificar en primera instancia por el tamaño de los clastos que presentan: • Conglomerados, cla
21	ente, los procesos diagenéticos cementarán los clastos resultantes de la erosión y el transporte
22	astos entre 2 mm y 0,06 mm. • Lutitas, si los clastos son menores de 0,06 mm; si los tamañ
23	o-médicas de Linares-huero". Está constituida por bioclastos (equinidos, briozoos, foraminíferos,
24	bioesparita, constituida en más de un 50% por bioclastos y con proporción Esparita/Micrita d
25) placa, 137 del Museo de Ciencias presentan intraclastos de gran tamaño (1 a 5 cm) y en alg
26	calclititas, ya que en casi todos los casos son exoclastos). Conglomerados (5,4%): marmoriza
27	erados (fig.2), placa 131 del M. de Ciencias son clasto- soportados y con frecuentes contactos
28	acterizando principalmente la morfología de sus clastos y la composición mineralógica de los s
29	nititas, tobas pozolánicas, coladas basálticas y piroclastos soldados constituyen los principales

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

	Word 1	Freq	Word 2	Freq	Text	Gap	Joint	MI	Z	MI3	Log L
5935	CALIZA	249	ORNAMENTAL	137	4	1	7	6.93	8.49	-9.17	32.97
5939	CALIZA	249	GRIS	179	4	1	7	6.55	7.22	-9.17	11.5
5945	CALIZA	249	AZUL	44	1	1	6	8.35	13.55	-9.84	158.3
5947	CALIZA	249	COMPACTA	50	5	1	7	8.39	14.84	-9.17	144.53
5951	CALIZA	249	PACKSTONE	5	1	1	4	10.9	27.5	-11.6	302.94
5952	CALIZA	249	CRISTALINA	50	3	1	3	7.16	6.1	-12.84	144.53
5954	CALIZA	249	BLANCA	34	3	1	7	8.94	18.17	-9.17	184.48
5956	CALIZA	249	BLANDA	22	3	1	3	8.35	9.58	-12.84	223.03
5962	CALIZA	249	FOSILÍFERA	8	2	1	4	10.22	21.67	-11.6	285.02
5963	CALIZA	249	DOLOMITIZADA	5	3	1	5	11.22	34.41	-10.63	302.94
5967	CALIZA	249	DOLOMÍTICA	9	1	1	4	10.05	20.41	-11.6	279.58
5968	CALIZA	249	BRECHOIDE	15	1	1	4	9.32	15.71	-11.6	250.81
5971	CALIZA	249	ARRECFAL	5	1	1	5	11.22	34.41	-10.63	302.94
5972	CALIZA	249	MARMOREA	11	7	1	8	10.76	37.06	-8.6	269.33
5973	CALIZA	249	PORTLAND	21	1	1	4	8.83	13.19	-11.6	226.71
5975	CALIZA	249	FRANCESA	10	2	1	5	10.22	24.23	-10.63	274.36
5976	CALIZA	249	IRLANDESA	6	1	1	3	10.22	18.77	-12.84	296.65
5977	CALIZA	249	CRETÁCICA	9	1	1	5	10.37	25.56	-10.63	279.58
5984	CALIZAS	314	ORNAMENTALES	253	11	1	38	8.15	31.72	-1.85	6.58
6000	CALIZAS	314	GRIS	179	2	1	3	4.99	2.11	-12.84	37.44
6007	CALIZAS	314	COMPACTAS	31	5	1	7	8.74	16.9	-9.17	269.75
6012	CALIZAS	314	DOLOMÍTICAS	13	3	1	3	8.77	11.19	-12.84	343.99
6013	CALIZAS	314	LACUSTRES	10	1	1	4	9.57	17.18	-11.6	359.91
6014	CALIZAS	314	BIOCLÁSTICAS	5	3	1	3	10.15	18.3	-12.84	390.75
6015	CALIZAS	314	MARMÓREAS	20	8	1	19	10.81	58.14	-4.85	311.63
6016	CALIZAS	314	RECRISTALIZADAS	8	2	1	5	10.21	24.12	-10.63	371.47
6021	CALIZAS	314	ARENOSAS	9	1	1	6	10.3	27.3	-9.84	365.58
6022	CALIZAS	314	FOSILÍFERAS	8	3	1	7	10.69	33.85	-9.17	371.47
6026	CALIZAS	314	OOLÍTICAS	5	3	1	5	10.89	30.61	-10.63	390.75
6027	CALIZAS	314	CRETÁCIAS	10	4	1	9	10.74	38.93	-8.09	359.91
6028	CALIZAS	314	JURÁSICAS	8	2	1	6	10.47	28.98	-9.84	371.47
6029	CALIZAS	314	CONTINENTALES	7	2	1	4	10.08	20.61	-11.6	377.6
6030	CALIZAS	314	TERCIARIAS	12	2	1	8	10.3	31.53	-8.6	349.13
6031	CALIZAS	314	MICRÍTICAS	8	2	1	4	9.89	19.25	-11.6	371.47
6032	CALIZAS	314	MARMORIZADAS	9	1	1	7	10.52	31.89	-9.17	365.58

The content: TMS in Translation Memories



The content – TMS in Translation Memories

- ❖ A TMS can help with various aspects of the translator's terminology-related tasks, including:
 - Storage
 - Retrieval
 - Updating of term entries.
- ❖ Today all Translation Memory Systems have integrated terminology management tools

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – TMS in Translation Memories

The screenshot displays a software interface with a grid on the left and two main text panels. The top panel shows the English term 'mallet' with its part of speech (n), subject (TOOL), and various definitions and context references. The bottom panel shows the Spanish term 'mazo' with its grammatical category (n), conceptual field (HERRMIA), and similar definitions. A status bar at the bottom indicates the search path: 'S: mallet F: Inglés (Reino Unido) -> Esp: Stone (NaturalStone)'. A URL 'Chelo Vargas - http://www.ua.es/personal/chelo.vargas/index.html' is visible at the bottom right.

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – TMS in Translation Memories

The screenshot displays a software interface with a grid on the left and two main text panels. The top panel shows the English term 'vener' with its grammatical category (CONST) and a detailed definition: 'The stone is set as a nonload bearing veneer with horizontal support at each floor and appropriate anchors to the structural wall behind aplacado chapa, lámina'. The bottom panel shows the Spanish term 'vener' with its grammatical category (CONST) and a definition: 'an outside, non-bearing load wythe of masonry used as a facing material'. An image of a stone wall is included in the bottom panel. A status bar at the bottom indicates the search path: 'S: F: Inglés (Reino Unido) -> Esp: Stone (NaturalStone)'. A URL 'Chelo Vargas - http://www.ua.es/personal/chelo.vargas/index.html' is visible at the bottom right.

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – TMS in Translation Memories

The screenshot displays the Transit XY software interface. It features a menu bar (Archivo, Edición, Cursor, Ver, Proyecto, Opciones, Ventana, Ayuda) and a toolbar. The main workspace is divided into several panes:

- English (United States): tutorial**: Shows a source text snippet: «14+»Transit is a comprehensive and sophisticated computer-aided translation system.«15».
- Spanish: Tutorial**: Shows the corresponding target text: «14+»Transit es un sistema de traducción asistida por ordenador «15».
- Project Management**: A sidebar containing project settings such as «18+»Inheriting project settings«19», «20+»Alias names«21», «22+»Detailed status information and extensive statistics«23», and «24+»Easier, yet more powerful segmentation«25».
- Dictionary: tutorial**: A pane at the bottom showing a definition for 'computer-aided translation system: sistema de traducción asistida por ordenador [A] [Tutorial project, Transit XY/Tutoria (Terminología)]'. This pane is circled in red.

At the bottom of the window, the text 'Chelo Vargas - <http://www.ua.es/personal/chelo.vargas/index.html>' is visible.

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content – TMS in Translation Memories

The screenshot displays the Deja Vu X software interface. The main window is titled 'C:\ProyectoDVX\Impresión.dvtdb'. It features a menu bar (Archivo, Editar, Ver, Base de datos, Usuarios, Herramientas, Ventana, Ayuda) and a toolbar. The interface is divided into several panes:

- Idioma de origen:** Inglés (Estados Unidos)
- Localizar término:** Ink cartridge
- Idioma de destino:** Español
- Related Terms:** Cartridge de tinta negra
- Attributes:** Part of Speech: Noun, Gender: Masculine, Number: Singular, Subject: 0043 - Computer hardware, Client: 13 - Microsoft
- Attributes (right):** Subject: 0043 - Computer hardware, Client: 13 - Microsoft, Context: 0026206-(5)-Cartridge de tinta negra

At the bottom of the window, the text 'c.html' is visible.

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The content: KEY POINTS

- ❖ Contributions from computer science have implied changes in the methodology of terminology and term processing.
- ❖ There are 5 basic points in terminography at which computers play a highly significant role.
- ❖ When designing a corpus, the user should determine the precise criteria by analysing the goals of his/her project.
- ❖ The process of term extraction is computer-aided than fully automatic.
- ❖ There are 3 main approaches to term extraction (linguistic, statistical, hybrid).
- ❖ TMS can help translators with terminology-related tasks.
- ❖ TMS are included in all Translation MS.

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

Methods and Software for the Extraction, Manipulation and Management of Bilingual Terminology Data

The contents – References

- Bowker, L. y Pearson, J. (2002): *Working with Specialized Language*. London & New York: Routledge
- Cabré, M.T. (1998): *Terminology. Theory, methods and applications*. Amsterdam/Philadelphia: John Benjamins..
- Cabré, M.T., Estopà, R. y Vivaldi, J. (2001): «Automatic term detection. A review of current systems», en Bourigault, D., Jacquemin, C. y L'Homme, M.C. (eds.): *Recent Advances in Computational Terminology*, Amsterdam/Philadelphia: John Benjamins, pp. 54-87.
- Estopà, R. (1999): *Extracció de terminologia: elements per a la construcció d'un SEACUSE*, Tesis Doctoral, Barcelona: Institut Universitari de Lingüística Aplicada. Disponible en línia: <http://www.tdx.cesca.es/>

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