The GEROM Project Search Engine

Marco Stefano Tomatis
Università degli Studi Internazionali di Roma - UNINT

Abstract

This paper aims at describing the technical engine underlying the project "GEROM", an online bilingual (German-Italian) terminological tool which has been developed by the Johannes Gutenberg - Universität of Mainz/Germersheim, Germany and Unint - University of International Studies of Rome, Italy. GEROM has been designed to facilitate the comprehension, analysis and translation of Italian and German texts dealing with different social and cultural aspects of the two Countries which are taken into account. Therefore, in order to provide users with a terminological support which focuses on examples of real language, a corpus-based approach has been adopted. Furthermore, since the core of the system is founded on a web-oriented service, a robust search interface based on CGI (Common Gateway Interface) technology has been implemented.

Key Words: terminology; CGI; search engine

Intercultural communication has always represented a complex, multifaceted issue. For this reason, it has been dealt in different ways and with different approaches by linguistic science in its broadest sense. Whereas from a theoretical point of view the gap between signifier and meaning, which represents the core of the problem, has been managed by various disciplines ranging from logics and semantics to philosophy of language, its practical evidence could be effectively faced only after the computer science provided linguists with a robust technological framework. The continuous improving of processing speed, storage space and programming languages which make pattern matching of linguistic structures an easy task provided new impulse in linguistic research, giving birth to totally new disciplines like Computational Linguistics, Corpus Linguistics, and Machine Translation.
Amongst the above mentioned research fields, Corpus Linguistics is, in absolute terms, the most exploited as it is the only one that can provide scholars and researchers with data mirroring the real usage of a language. The Gerom project is developed starting from this corpus-driven approach. It aims at improve the translation of German and Italian texts by showing the real usage of language-specific terminology in a variety of contexts.

1. Linguistic data management

In order to reduce the system structure as much as possible, all linguistic data have been stored in a web server in plain text, UTF-8 format and no server-side database or data management software has been adopted. Such a choice, which may sound strange to most Information Technology professionals, has been made on purpose. Indeed, despite their renowned efficiency, in general databases require specific server technology to achieve good response performances when many concurrent search queries are entered.

Furthermore, databases require periodical maintenance sessions and serious problems may occur if this task is not regularly accomplished. Another important issue is tied to data management aspects. For security reasons and retrieval effectiveness, all databases convert plain text into an indexed, binary-coded structure which cannot be easily accessed in real time by other software products. Data exchange is possible only after a specific table export process is run on demand.

As a consequence of this, the human operator is, in fact, impeded to visualize and modify the linguistic information with a simple text editor. On the contrary, all the tables which are stored in the server can be accessed and manipulated via SQL language by the same database management system which has been used to create them. If, on the one hand, this feature may greatly improve the search engine efficiency, on the other hand it seriously limits the data structure flexibility, making changes hard to implement in case sudden and relevant modifications of the linguistic knowledge-base are required (i.e. changing the primary index or adding a field or a record in the middle of a table in use).

1.1 Information structure

Although a complete and thorough description of the linguistic dataset which has been adopted goes beyond the scope of this paper, some words will be spent on this subject. Firstly, it is important to keep in mind that the linguistic resources the system may refer to, cover two functions and differentiate accordingly. As for the role, they play the primary and most important resource deals with providing users with all the relevant terminological information tied to any German or Italian search query which is entered in the GEROM home page. To accomplish this service, the content of a particular file which works like an enriched bilingual terminological dictionary is read.

Structurally organized like a spreadsheet, this file is a tab-separated value UTF-8 plain text in which each record contains 75 different fields ranging from the terminological expression and its synonyms, to the name of the corpora where the said expression can be found in. Because of its leading function, the bilingual terminological resource is the first file the CGI engine must access. Only after the user's request has exactly matched the content of one specific field of the said terminological file, all the query-related linguistic information is extracted and properly mark upped to be displayed in HTML format. The second kind of linguistic resources, instead, is represented by all the different corpora available in the server. Each corpus is an XML-structured collection of plain text, single theme, monolingual articles which are enriched with relevant metadata information. As a final note, it is useful to remark that all the linguistic resources the system may use have been edited by Professor Claudio Fantinuoli of the Johannes Gutenberg University and Professor Fabio Proia of Unint.

2. *CGI scripts*

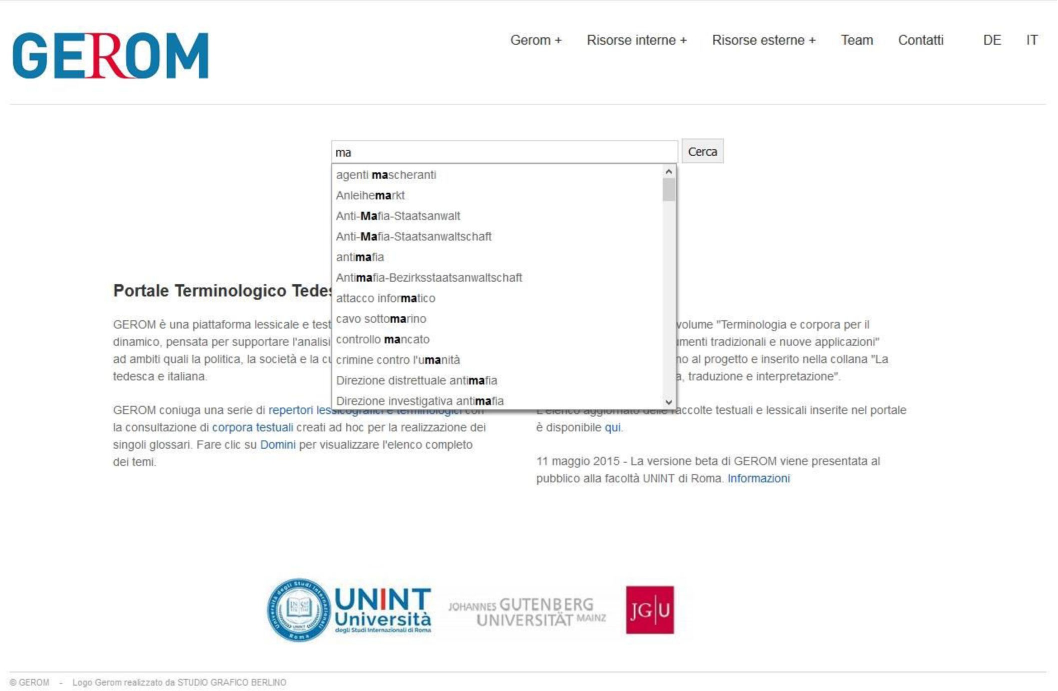
As mentioned before, the interaction between the user requests and the web server is managed by CGI technology. Without facing specific technical details, the most important feature characterizing this approach is the possibility to establish a dynamic client-server communication by adopting any suitable server-side programming language. Since the system is required to manipulate plain text strings only, the most convenient way to do this is choosing a text-oriented programming language like AWK, Perl, or the most recent Python. Despite their differences in terms of specific scope and purpose they were designed for, all of the above-mentioned computer languages descend from "C" and offer a powerful backward-compatible, regexp-centered, high level, straightforward syntax. These features are of primary importance because they allow developers to easily maintain or upgrade their program by adding new functions at any time or, if needed, translate the source code from a scripting language to another with minimal adaptations.

In order to achieve a better visualization of the terminological information and improve the interactive exploitation of the online platform, two different CGI algorithms have been created. The most complex of them has been designed to manage multiple functions, so it represents the core of the system. The other one, instead, plays a secondary role. It activates on user demand only and limits to extract and display a wider area of the text the main program returns to the general response page.

2.1 Input query

For clarity reasons, the CGI script which is run for first will be described by steps, focusing on its functional workflow. As mentioned before, this program performs a number of different tasks which range from dealing with all the incoming requests from remote users, to the creation of all the response pages. Therefore, this script is automatically run by the Apache web server whenever a query is entered in the GEROM home page search box. Because the first task of the script is designed to handle the user queries, particular features of the HTTP protocol should be properly managed in order to obtain an orthographically standard input string.

Although apparently strange, this operation is mandatory because, according to the transfer protocol, any query can be transmitted to the web server in plain ASCII code only. This implies that all the accented letters and the characters with diacritical marks are automatically transformed in a non-intelligible string before being sent to the server (e.g. "à" > "%C3%A0"). Starting from this assumption, it becomes clear that in order to avoid mismatch issues, the whole user query must undergo a reconversion process before being stored in a variable for any further use. As regards queries, it is important to put in evidence that the system has been designed to accept both single and multiword terms as an input string. Moreover, to help users avoid any possible query misspelling error, a pre-compiled, close list of searchable terms has been made available in the home page search box (see Figure 1).


GEROM
Gerom +
Risorse interne +
Risorse esterne +
Team
Contatti
DE
IT
Cerca
UNINT Università degli Studi Internazionali di Roma
JOHANNES GUTENBERG UNIVERSITÄT MAINZ
JG|U

ma

agenti **ma**scheranti
Anleihe**ma**rkt
Anti-**Ma**fia-Staatsanwalt
Anti-**Ma**fia-Staatsanwaltschaft
anti**ma**fia
Anti**ma**fia-Bezirksstaatsanwaltschaft
attacco infor**ma**tico
cavo sotto**ma**rino
controllo **ma**ncato
crimine contro l'u**ma**nità
Direzione distrettuale anti**ma**fia
Direzione investigativa anti**ma**fia

Portale Terminologico Tedes

GEROM è una piattaforma lessicale e test
dinamico, pensata per supportare l'analisi
ad ambiti quali la politica, la società e la cu
tedesca e italiana.

GEROM coniuga una serie di repertori les
la consultazione di corpora testuali creati ad hoc per la realizzazione dei singoli glossari. Fare clic su Domini per visualizzare l'elenco completo dei temi.

volume "Terminologia e corpora per il
menti tradizionali e nuove applicazioni"
no al progetto e inserito nella collana "La
a, traduzione e interpretazione".

accolte testuali e lessicali inserite nel portale è disponibile qui.

11 maggio 2015 - La versione beta di GEROM viene presentata al pubblico alla facoltà UNINT di Roma. Informazioni

© GEROM - Logo Gerom realizzato da STUDIO GRAFICO BERLINO

Figure 1: Home page search box

2.2 Response page

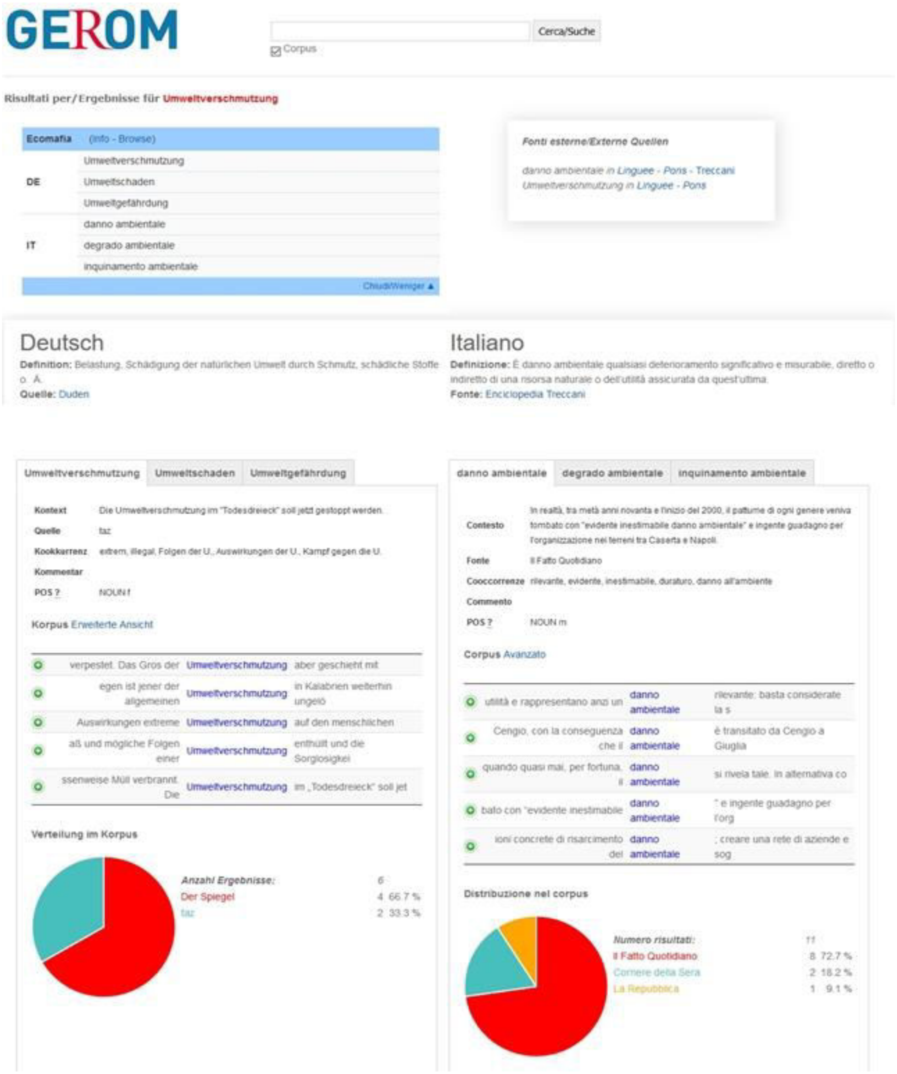
The next functional block of the CGI program is designed to produce the response pages to the clients by extracting all the required information from both the terminological knowledge-base and the corpora. To carry out this complex task, the first step must consist in printing the header lines which are required to create a UTF-8 encoded HTML page. In addition, all the Javascript and CSS codes which have been edited by Claudio Fantinuoli to enrich the web page with graphical features are included in this section, too.

After the header block declaration has been set, the next step regards selecting and retrieving text samples from the files which are stored in the web server. To do this, the program iteratively scans all the lines the terminological file is made of to find any possible match with the user's query string. All pattern matching tests are managed by regular expressions, which provide the possibility to define exact or partial matching rules. In our case, any pattern matching test will pass only if the user input string and a specific field of the terminological file are proven to be identical. After a correct match is found, the whole set of metadata, flexional alternatives and homonyms tied to the query string are saved in dedicated variables and an array is set to store the links to the corpora which are eligible to contain the text patterns to find.

2.3 Corpus-driven approach

After collecting all the useful data from the terminological reference file, the program uses the same pattern matching technique to sample a limited number of lines from the subject-related corpora. This step is very important because it allows the system to provide users with a short list of examples which represent the basic stage to understand the real use of language in its own specific context. Yet, differently from the operations involving the terminological dataset, the lookup and text extraction activity made on corpora is driven by both the query string and its own related homonyms and flexional variants, too.

Summing up, the whole text mining process implements two nested loops. The former manages the pattern matching between the input term and all the lines the terminological file is made of; the latter matches the content of the selected corpora with both the term to search and, if any, its semantic and morphologic variants. After concluding the pattern matching activity on one corpus, the same processing is run on the corpus which contains the corresponding terminological information in the other language GEROM offers. During the above mentioned tasks, pertinent linguistic data is retrieved, collected and properly organized in an html page that will be transmitted to the user web browser as the end of the terminological file is met (see Figure 2).



Risultati per/Ergebnisse für Umweltverschmutzung

Ecomafia	(Info - Browse)
DE	Umweltverschmutzung Umweltschaden Umweltgefährdung
IT	danno ambientale degrado ambientale inquinamento ambientale

Fonti esterne/Externe Quellen

danno ambientale in Linguee - Pons - Treccani
Umweltverschmutzung in Linguee - Pons

Deutsch

Definition: Belastung, Schädigung der natürlichen Umwelt durch Schmutz, schädliche Stoffe o. Ä.
Quelle: Duden

Italiano

Definizione: È danno ambientale qualsiasi deterioramento significativo e misurabile, diretto o indiretto di una risorsa naturale o dell'utilità assicurata da quest'ultima.
Fonte: Enciclopedia Treccani

Kontext	Die Umweltverschmutzung im "Todesdreieck" soll jetzt gestoppt werden.
Quelle	taz
Kookkurrenz	extrem, illegal, Folgen der U., Auswirkungen der U., Kampf gegen die U.
Kommentar	
POS ?	NOUN f

Korpus Erweiterte Ansicht

⊕	verpestet. Das Gros der	Umweltverschmutzung	aber geschieht mit
⊕	egen ist jener der allgemeinen	Umweltverschmutzung	in Kalabrien weiterhin ungelö
⊕	Auswirkungen extreme	Umweltverschmutzung	auf den menschlichen
⊕	aß und mögliche Folgen einer	Umweltverschmutzung	enthüllt und die Sorglosigkei
⊕	ssenweise Müll verbrannt. Die	Umweltverschmutzung	im „Todesdreieck" soll jet

Verteilung im Korpus

Anzahl Ergebnisse:	6
Der Spiegel	4 66.7 %
taz	2 33.3 %

Contesto	In realtà, tra metà anni novanta e l'inizio del 2000, il pattume di ogni genere veniva tombato con "evidente inestimabile danno ambientale" e ingente guadagno per l'organizzazione nei terreni tra Caserta e Napoli.
Fonte	Il Fatto Quotidiano
Cooccorrenze	rilevante, evidente, inestimabile, duraturo, danno all'ambiente
Commento	
POS ?	NOUN m

Corpus Avanzato

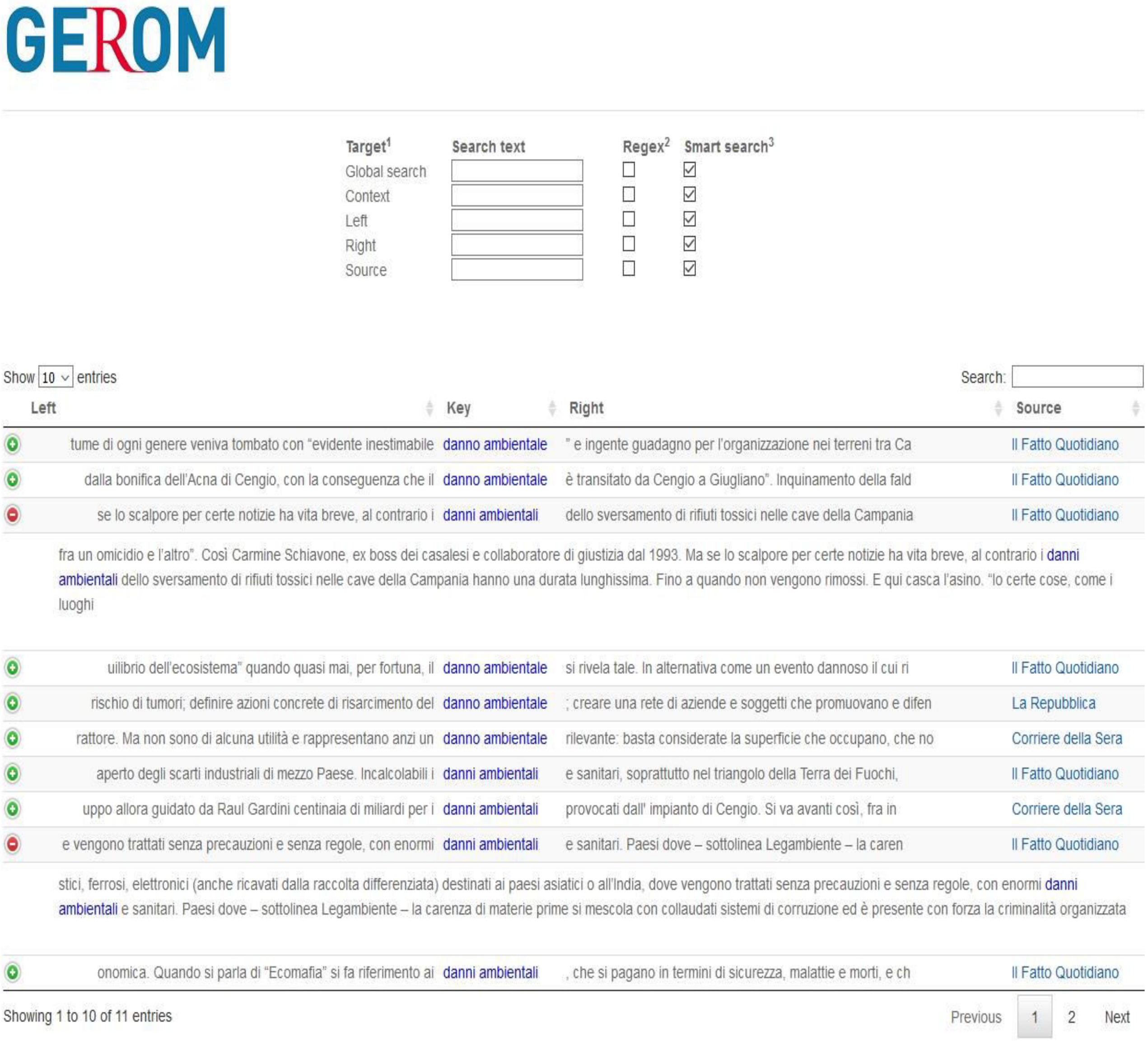
⊕	utilità e rappresentano anzi un	danno ambientale	rilevante: basta considerate la s
⊕	Cengio, con la conseguenza che il	danno ambientale	è transitato da Cengio a Giuglia
⊕	quando quasi mai, per fortuna, il	danno ambientale	si rivela tale. In alternativa co
⊕	bato con "evidente inestimabile	danno ambientale	" e ingente guadagno per l'org
⊕	ioni concrete di risarcimento del	danno ambientale	; creare una rete di aziende e sog

Distribuzione nel corpus

Numero risultati:	11
Il Fatto Quotidiano	8 72.7 %
Corriere della Sera	2 18.2 %
La Repubblica	1 9.1 %

Figure 2: Basic response page

Another important feature offered by GEROM regards the possibility to overcome the reduced number of elements the basic response page is designed to return and enhance the visualization of linguistic information by providing users with a wider range of concordance lines extracted from the corpora in use. To achieve this goal, a specific link to a dedicated CGI script has been set in each tab the user response page holds. For the script be able to work, relevant parameters like the name of the corpus where information should be retrieved from and the term to find - the original query string or one of its linguistic variants - must be transmitted. Like before, also in this case a HTML page is returned to the user, but it will be opened in a new window for readability reasons (see Figure 3). As for the text lines which are sampled from the corpora to be included into the response page, it is worth noting that the adoption of a dedicated hypertext markup structure was mandatory to display the whole data in a KWIC (Key Word In Context) compliant format.



GEROM

Target[1]	Search text	Regex[2]	Smart search[3]
Global search		□	☑
Context		□	☑
Left		□	☑
Right		□	☑
Source		□	☑

Show 10 entries Search:

	Left	Key	Right	Source
⊕	tume di ogni genere veniva tombato con "evidente inestimabile	danno ambientale	" e ingente guadagno per l'organizzazione nei terreni tra Ca	Il Fatto Quotidiano
⊕	dalla bonifica dell'Acna di Cengio, con la conseguenza che il	danno ambientale	è transitato da Cengio a Giugliano". Inquinamento della fald	Il Fatto Quotidiano
⊖	se lo scalpore per certe notizie ha vita breve, al contrario i	danni ambientali	dello sversamento di rifiuti tossici nelle cave della Campania	Il Fatto Quotidiano
	fra un omicidio e l'altro". Così Carmine Schiavone, ex boss dei casalesi e collaboratore di giustizia dal 1993. Ma se lo scalpore per certe notizie ha vita breve, al contrario i danni ambientali dello sversamento di rifiuti tossici nelle cave della Campania hanno una durata lunghissima. Fino a quando non vengono rimossi. E qui casca l'asino. "Io certe cose, come i luoghi			
⊕	uilibrio dell'ecosistema" quando quasi mai, per fortuna, il	danno ambientale	si rivela tale. In alternativa come un evento dannoso il cui ri	Il Fatto Quotidiano
⊕	rischio di tumori; definire azioni concrete di risarcimento del	danno ambientale	; creare una rete di aziende e soggetti che promuovano e difen	La Repubblica
⊕	rattore. Ma non sono di alcuna utilità e rappresentano anzi un	danno ambientale	rilevante: basta considerate la superficie che occupano, che no	Corriere della Sera
⊕	aperto degli scarti industriali di mezzo Paese. Incalcolabili i	danni ambientali	e sanitari, soprattutto nel triangolo della Terra dei Fuochi,	Il Fatto Quotidiano
⊕	uppo allora guidato da Raul Gardini centinaia di miliardi per i	danni ambientali	provocati dall' impianto di Cengio. Si va avanti così, fra in	Corriere della Sera
⊖	e vengono trattati senza precauzioni e senza regole, con enormi	danni ambientali	e sanitari. Paesi dove – sottolinea Legambiente – la caren	Il Fatto Quotidiano
	stici, ferrosi, elettronici (anche ricavati dalla raccolta differenziata) destinati ai paesi asiatici o all'India, dove vengono trattati senza precauzioni e senza regole, con enormi danni ambientali e sanitari. Paesi dove – sottolinea Legambiente – la carenza di materie prime si mescola con collaudati sistemi di corruzione ed è presente con forza la criminalità organizzata			
⊕	onomica. Quando si parla di "Ecomafia" si fa riferimento ai	danni ambientali	, che si pagano in termini di sicurezza, malattie e morti, e ch	Il Fatto Quotidiano

Showing 1 to 10 of 11 entries

Figure 3: Enhanced response page

3. Future development

Although the basic framework the GEROM web platform is founded on has achieved a stable operating structure, for some aspects the project should be still considered as a work-in-progress. In particular, further development are planned to increase the linguistic resources offered by the system in terms of both subject-specific corpora and external references users may get access to. Gerom is freely accessible at the following web address: http://gerom.unint.eu/it/index.html

References

Zhang, C.; Wu, D. (2012). Bilingual Terminology Extraction Using Multi-level Termhood. In: *The Electronic Library*, 30(2), pp. 295-309

Fantinuoli, C.; Proia, F. (eds.). (2015). *Terminologia e corpora per il dialogo italo-tedesco. Strumenti tradizionali e nuove applicazioni.* Roma, Aracne.

Robbins, A. (2015). *Gawk: Effective AWK Programming, 4th Edition.* Universal Text Processing and Pattern Matching. Sebastopol, O'Reilly Media.

Scott, G.; Shishir, G.; Gunther, B. (1999). *CGI Programming with Perl, 2nd Edition.* Sebastopol, O'Reilly Media.

Jeffrey, E.F.F. (2006). *Mastering Regular Expressions, 3rd Edition.* Sebastopol, O'Reilly Media.

Mari-Carmen Marcos et al. (2006). Usability evaluation of online terminology databases. In: *Hipertext.net*, 4. Accessed at: http://www.hipertext.net [19/10/2015]

Zengin, B. & Kacar, I.G. (2011). Turkish EFL Academicians' Problems Concerning Translation Activities and Practices, Attitudes towards the Use of Online and Printed Translation Tools, and Suggestions for Quality Translation Practice. In: *Turkish Online Journal of Educational Technology*, 10(2), 274-286.

Ahmet Aker, Monica Lestari Paramita, Marcis Pinnis (Tilde), Robert Gaizauskas. 2014. Bilingual dictionaries for all EU languages. *Proceedings of the Ninth International Conference on Language Resources and Evaluatio*n (LREC'14), 2839-2845.

Tatiana Gornostay and Andrejs Vasiļjevs. 2014. Cloud Terminology Services Facilitate Specialised Lexicography Work. In *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 15-19 July 2014, Bolzano/Bozen, pp. 621-629.

Vasiljevs, A., Pinnis, M., & Gornostay, T. (2014). Service model for semi-automatic generation of multilingual terminology resources. In: *Terminology and Knowledge Engineering 2014*. Accessed at: https://hal.archives-ouvertes.fr/hal-01005874 [19/10/2015]