

A Controlled Language Checker based on the Ellogon text Engineering Platform

Vangelis Karkaletsis, Georgios Samaritakis, Georgios Petasis, Dimitra Farmakiotou,
Ion Androutsopoulos and Constantine D. Spyropoulos

Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research (N.C.S.R.) "Demokritos"
P.O BOX 60228, Aghia Paraskevi, 153 10 Athens, Greece.
{vangelis, samarita, petasis, dfarmak, ionandr, costass}@iit.demokritos.gr

"Script Outline" of the demo presentation, including accompanying narrative

The checker's operation is presented using screen-shots accompanied with short descriptions. This presentation is performed at two levels:

- end-user level (i.e. the technical writer uses the checker as an extra operation included in his/her wordprocessor)
- expert user (i.e. the operations done internally)

End-user level

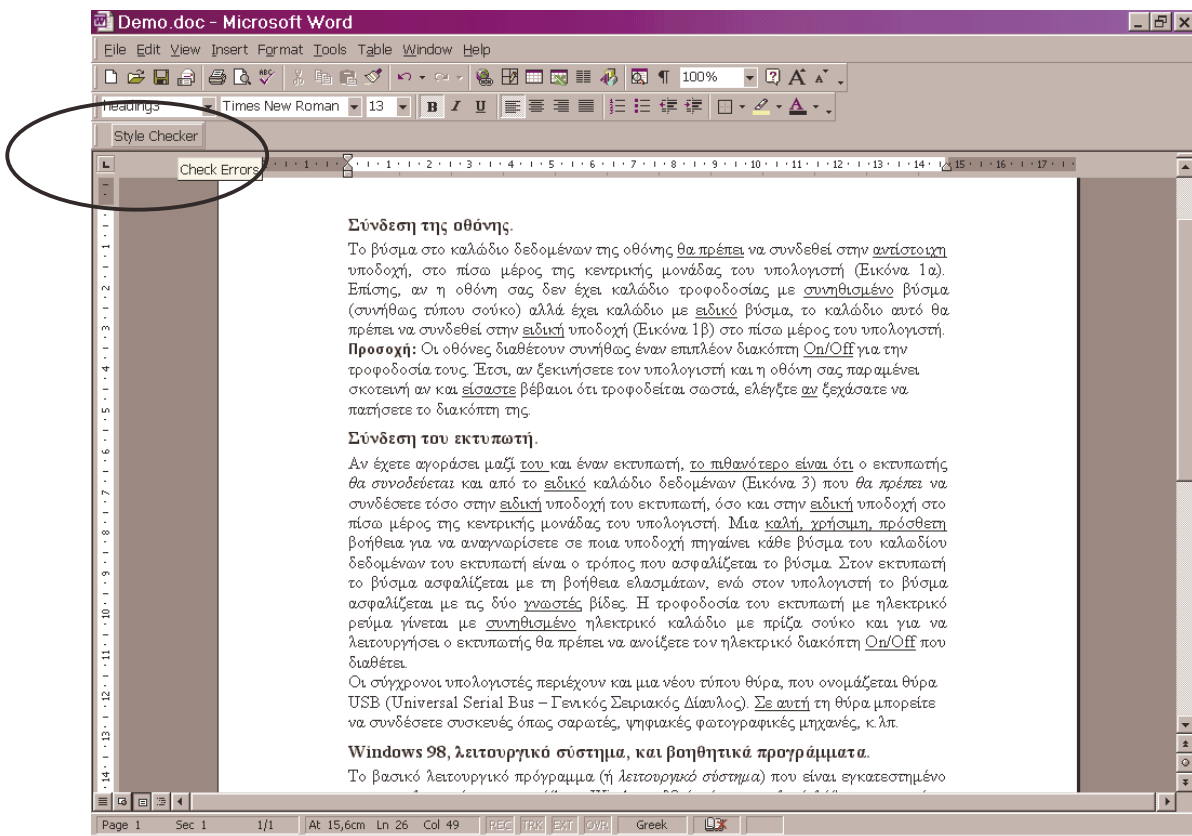


Fig.1 Call the style checker from the Wordprocessor: The user presses the "Style Checker" button

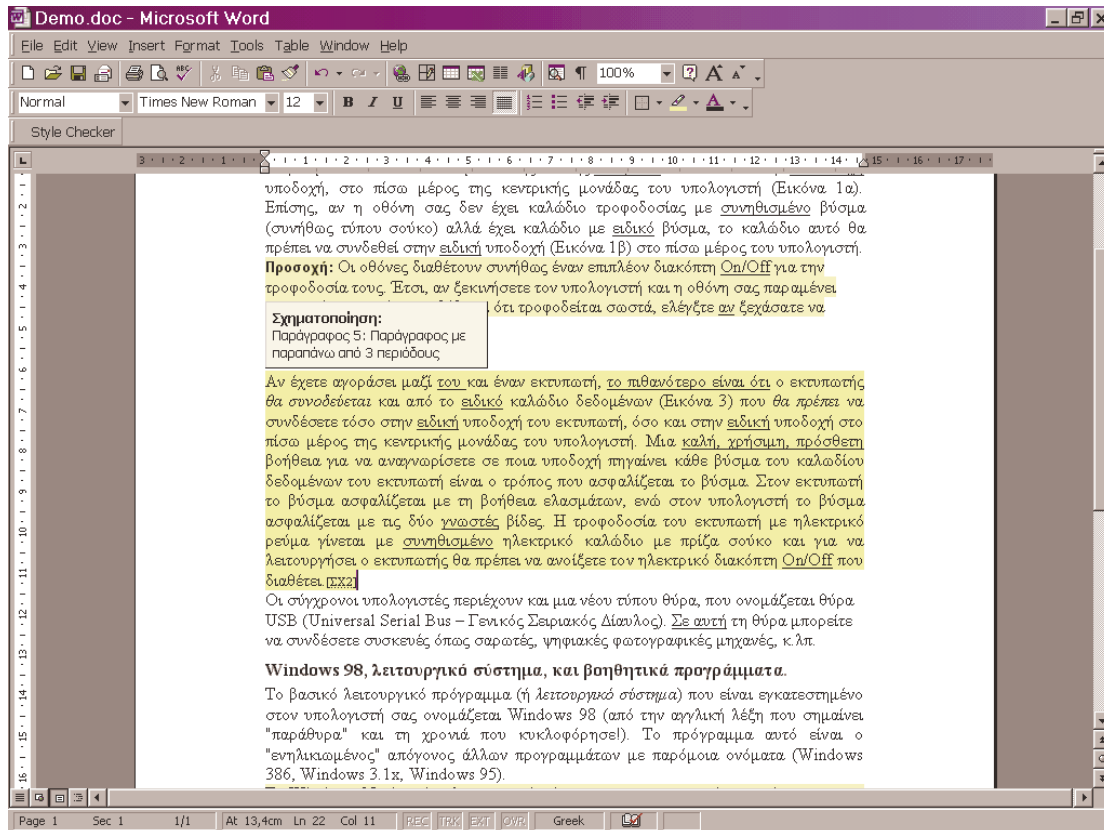


Fig.2 Errors at paragraph-level: The user sees the relevant comments

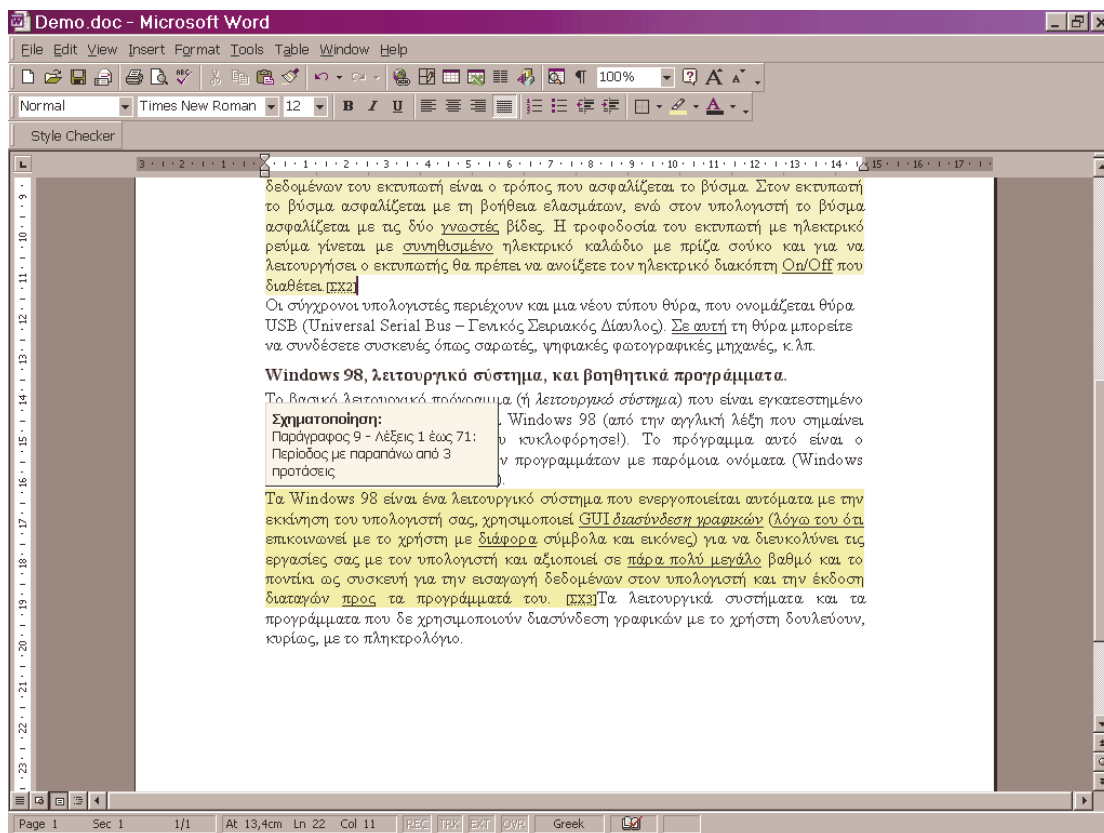


Fig.3 Errors at sentence-level: The user sees the relevant comments

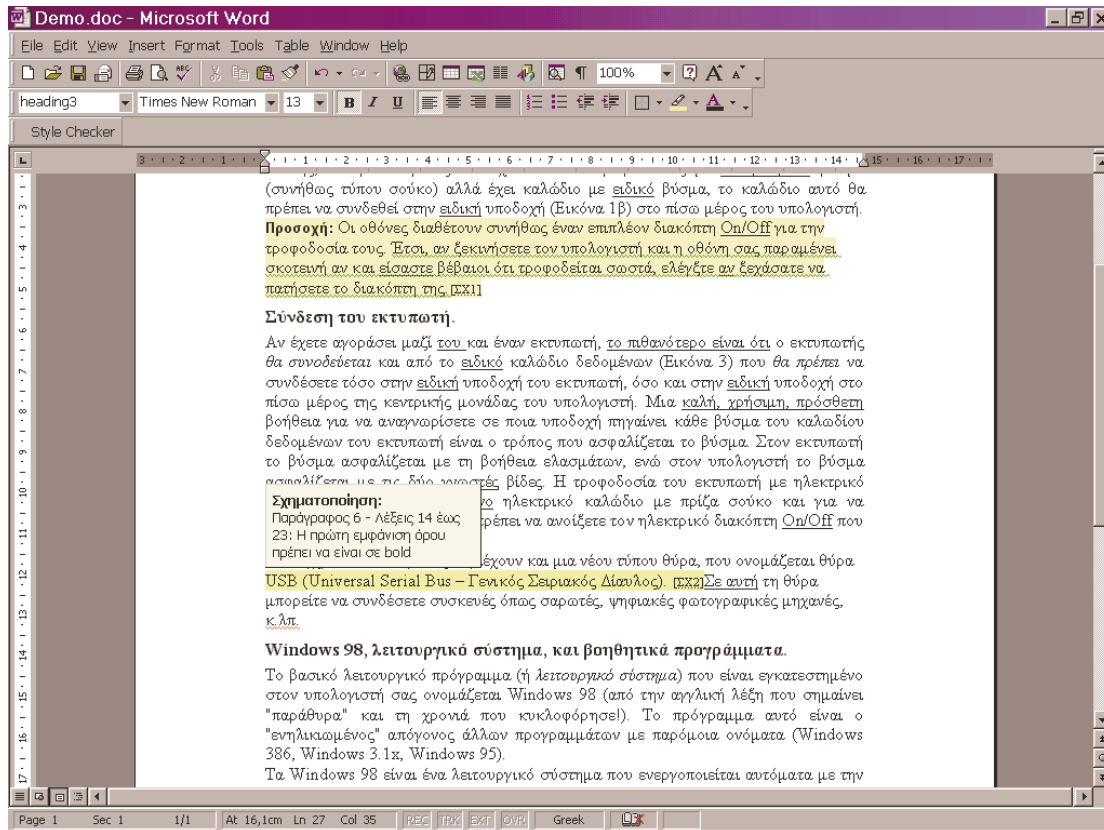


Fig.4 Errors at phrase-level: The user sees the relevant comments

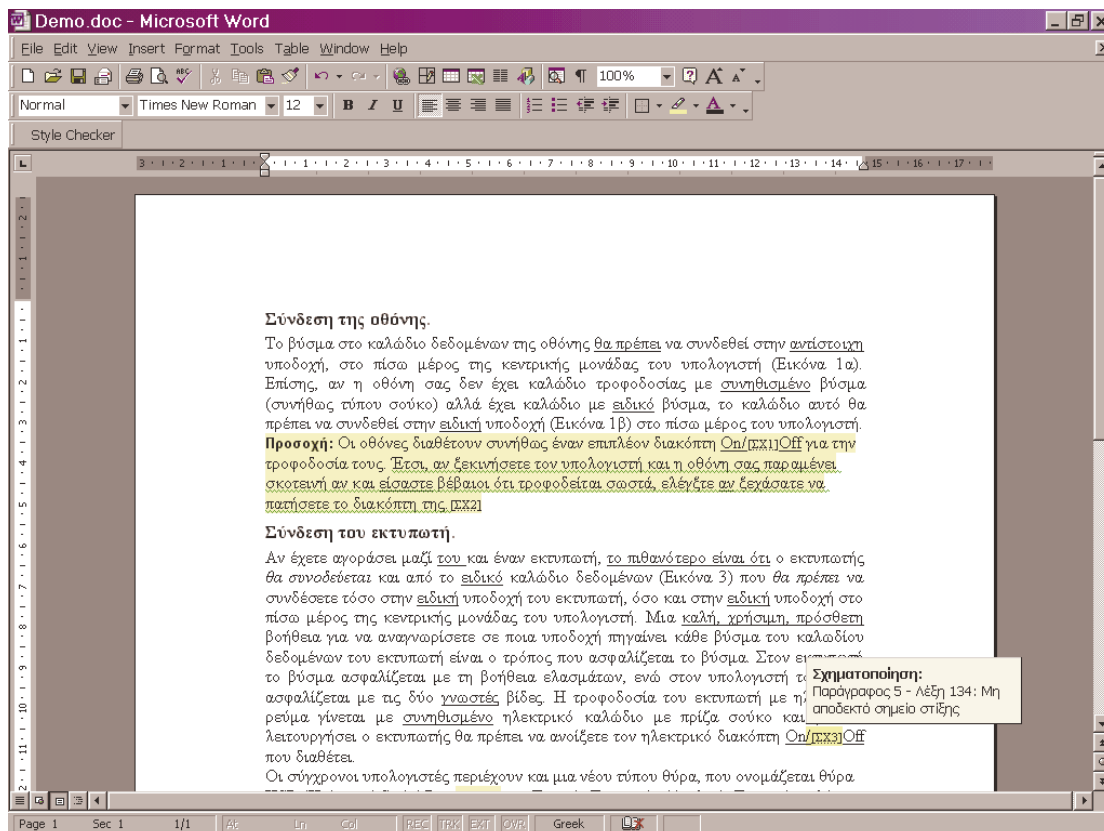


Fig.4 Errors at word-level: The user sees the relevant comments

Expert user level

```

<?xml version="1.0" encoding="ISO-8859-7"?>
<!DOCTYPE document SYSTEM 'morphi.dtd'>
<document>
<heading3 pid="1">
<text>
<word wid="1" style="Bold">Σύνδεση </word>
<word wid="2" style="Bold">της </word>
<word wid="3" style="Bold">οθόνης</word>
<word wid="4" style="Bold"></word>
</text>
</heading3>
<normal pid="2">
<text>
<word wid="1" style="Regular">Το </word>
<word wid="2" style="Regular">βύσμα </word>
<word wid="3" style="Regular">στο </word>
<word wid="4" style="Regular">καλώδιο </word>
<word wid="5" style="Regular">δεδομένων </word>
<word wid="6" style="Regular">της </word>
<word wid="7" style="Regular">οθόνης </word>
<word wid="8" style="Regular">θα </word>
<word wid="9" style="Regular">πρέπει </word>
<word wid="10" style="Regular">να </word>
<word wid="11" style="Regular">συνδεθεί </word>
<word wid="12" style="Regular">στην </word>
<word wid="13" style="Regular">αντίστοιχη </word>
<word wid="14" style="Regular">υποδοχή</word>
<word wid="15" style="Regular">, </word>
<word wid="16" style="Regular">στο </word>
<word wid="17" style="Regular">πίσω </word>
<word wid="18" style="Regular">μέρος </word>
<word wid="19" style="Regular">της </word>

```

Fig.5 The XML document (output of the XML generator)

ID	Type	Start	End	Attributes
0	heading3	0	18	pid=1
1	heading3	18	41	pid=2
2	normal	41	497	pid=3
3	normal	497	799	pid=4
4	heading3	799	841	pid=5
5	normal	841	1768	pid=6
6	normal	1768	2044	pid=7
7	heading3	2044	2113	pid=8
8	normal	2113	2504	pid=9
9	normal	2504	3161	pid=10

Annotations of Document Document1.txt, From Collection XML_demo

Κλείσιμο

Fig.6 Conversion to the Ellogon format: Ellogon shares the same data model as the TIPSTER architecture. The central element for storing data in Ellogon is the *Collection*. A collection is a finite set of *Documents*. An Ellogon document consists of textual data as well as linguistic information about the textual data. This linguistic information is stored in the form of *attributes* and *annotations*.

ID	Type	Start	End	Attributes
9	normal	2504	3161	pid=10
1365	sentence	0	17	constituents=[485 487 488]
1366	sentence	18	40	constituents=[490 492 494 495]
1367	sentence	41	220	constituents=[497 499 501 503]
1368	sentence	222	496	constituents=[548 549 551 553]
1369	sentence	497	604	constituents=[630 631 633 635]
1370	sentence	606	798	constituents=[661 662 664 666]
1371	sentence	799	840	constituents=[716 718 720 722]
1372	sentence	841	1183	constituents=[727 729 731 733]
1373	sentence	1185	1381	constituents=[827 829 830 832]

Σύνδεση της οθόνης .
 Το βύσμα στο καλώδιο δεδομένων της οθόνης θα πρέπει να συνδεθεί στην αντίστοιχη υποδοχή , στο πίσω μέρος της κεντρικής μονάδας του υπολογιστή (Εικόνα 1α) . Επίσης , αν η οθόνη σας δεν έχει καλώδιο τροφοδοσίας με συνηθισμένο βύσμα (συνήθως τύπου σούκο) αλλά έχει καλώδιο με ειδικό βύσμα , το καλώδιο αυτό θα πρέπει να συνδεθεί στην ειδική υποδοχή (Εικόνα 1β) στο πίσω μέρος του υπολογιστή .

Προσοχή : Οι οθόνες διαθέτουν συνήθως έναν επιπλέον διακόπτη On / Off για την τροφοδοσία τους . Έτσι , αν ξεκινήσετε τον υπολογιστή και η οθόνη σας

Κλείσιμο

Fig.8 Sentence splitter results

ID	Type	Start	End	Attributes
1113	token	2240	2247	type=EFW, pos=foreign
1114	token	2249	2251	type=NUMBER
1115	token	2253	2254	type=SYMBOL
1117	token	2255	2258	type=GLW, pos=AsPpSp
1119	token	2260	2263	type=GLW, pos=AtDf
1121	token	2265	2272	type=GLW, pos=AjFeSg
1123	token	2274	2278	type=GLW, pos=NoCmFeSg
1125	token	2280	2283	type=GLW, pos=PnRe
1127	token	2285	2293	type=GLW, pos=VbMnPr
1128	token	2295	2296	type=SYMBOL

Windows 98 , λειτουργικό σύστημα , και βοηθητικά προγράμματα .
 Το βασικό λειτουργικό πρόγραμμα (ή λειτουργικό σύστημα) που είναι εγκατεστημένο στον υπολογιστή σας ονομάζεται Windows 98 (από την αγγλική λέξη που σημαίνει " παράθυρα " και τη χρονιά που κυκλοφόρησε !) . Το πρόγραμμα αυτό είναι ο " ενηλικιωμένος " απόγονος άλλων προγραμμάτων με παρόμοια ονόματα (Windows 386 , Windows 3 . 1x , Windows 95) .
 Τα Windows 98 είναι ένα λειτουργικό σύστημα που ενεργοποιείται αυτόματα με την εκκίνηση του υπολογιστή σας , χρησιμοποιεί GUI διασύνδεση γραφικών (λόγω

Κλείσιμο

Fig.9 POS tagger results (use of PAROLE tags)

ID	Type	Start	End	Attributes
3745	morph	2154	2165	lemma=λειτουργικός, stem=λειτουργ, suffi
3746	morph	2154	2165	lemma=λειτουργικός, stem=λειτουργ, suffi
3747	morph	2154	2165	lemma=λειτουργικός, stem=λειτουργ, suffi
3748	morph	2167	2174	lemma=σύστημα, stem=σύστη, suffix=μα
3749	morph	2167	2174	lemma=σύστημα, stem=σύστη, suffix=μα
3750	morph	2167	2174	lemma=σύστημα, stem=σύστη, suffix=μα
3751	morph	2178	2181	lemma=που, stem=που, suffix=, syllab
3752	morph	2183	2188	lemma=είμαι, stem=, suffix=είναι, sy
3753	morph	2183	2188	lemma=είμαι, stem=, suffix=είναι, sy
3754	morph	2183	2188	lemma=είμαι, stem=, suffix=είναι, sy

Windows 98 , λειτουργικό σύστημα , και βοηθητικά προγράμματα .
 Το βασικό λειτουργικό πρόγραμμα (ή λειτουργικό σύστημα) που είναι εγκατεστημένο στον υπολογιστή σας ονομάζεται Windows 98 (από την αγγλική λέξη που σημαίνει " παράθυρα " και τη χρονιά που κυκλοφόρησε !). Το πρόγραμμα αυτό είναι ο " ενηλικιωμένος " απόγονος άλλων προγραμμάτων με παρόμοια ονόματα (Windows 386 , Windows 3 . 1x , Windows 95) .
 Τα Windows 98 είναι ένα λειτουργικό σύστημα που ενεργοποιείται αυτόματα με την εκκίνηση του υπολογιστή σας , χρησιμοποιεί GUI διασύνδεση γραφικών (λόγω του ότι επικοινωνεί με το χρήστη με διάφορα σύμβολα και εικόνες) για να διευκολύνει τις εργασίες σας με τον υπολογιστή και αξιοποιεί σε πάρα πολύ μεγάλο βαθμό και το ποντίκι ως συσκευή για την εισαγωγή δεδομένων στον υπολογιστή και την έκδοση διαταγών προς τα προγράμματά του . Τα λειτουργικά συστήματα και τα προγράμματα που δε χρησιμοποιούν διασύνδεση γραφικών με το χρήστη δουλεύουν , κυρίως , με το πληκτρολόγιο .

Κλείσιμο

Fig.10 Lemmatiser results

ID	Type	Start	End	Attributes
5024	morph	3146	3158	lemma=πληκτρολόγιο, stem=πληκτρ
5025	lookup	134	141	type=greek_terms, tag=
5026	lookup	436	443	type=greek_terms, tag=
5027	lookup	1064	1071	type=greek_terms, tag=
5028	lookup	1114	1121	type=greek_terms, tag=
5029	lookup	896	922	type=stop_phrases, tag=
5030	lookup	1261	1268	type=greek_terms, tag=
5031	lookup	3077	3113	type=greek_terms, tag=

Το πρόγραμμα αυτό είναι ο " ενηλικιωμένος " απόγονος άλλων προγραμμάτων με παρόμοια ονόματα (Windows 386 , Windows 3 . 1x , Windows 95) .
 Τα Windows 98 είναι ένα λειτουργικό σύστημα που ενεργοποιείται αυτόματα με την εκκίνηση του υπολογιστή σας , χρησιμοποιεί GUI διασύνδεση γραφικών (λόγω του ότι επικοινωνεί με το χρήστη με διάφορα σύμβολα και εικόνες) για να διευκολύνει τις εργασίες σας με τον υπολογιστή και αξιοποιεί σε πάρα πολύ μεγάλο βαθμό και το ποντίκι ως συσκευή για την εισαγωγή δεδομένων στον υπολογιστή και την έκδοση διαταγών προς τα προγράμματά του . Τα λειτουργικά συστήματα και τα προγράμματα που δε χρησιμοποιούν διασύνδεση γραφικών με το χρήστη δουλεύουν , κυρίως , με το πληκτρολόγιο .

Κλείσιμο

Fig.11 Lookup results: it identifies those words/phrases in the documents that are also included in the linguistic resources (terminology, vocabulary). It annotates the identified words/phrases with the relevant information (e.g. Greek term, English term, prohibited word/phrase)