Chapter 2. The Eurolect Observatory Multilingual Corpus

CHAPTER TWO

The *Eurolect Observatory Multilingual Corpus* Construction and query tools

Marco Stefano Tomatis
Università degli Studi Internazionali di Roma – UNINT
Università degli Studi di Torino

Abstract

This chapter aims to explain the corpus design of the *Eurolect Observatory Multilingual Corpus* and the steps required to build all the different monolingual corpora the project needed to accomplish its research objectives. The first two paragraphs after the general introduction will point out the differences and the overlaps that characterize all the corpora that the author of this paper was in charge of producing as a member of the UNINT research team and that were used in the *Eurolect Observatory Project* for text mining. After accurately defining the data collection and corpus building strategies adopted, this paper will describe the corpus search tool that was developed in order to help scholars look for and save samples of text from the whole corpus in a convenient and easy way.

Keywords: natural language processing; corpus linguistics; awk; corpus search tool; regular expressions; markup

1 Introduction

This chapter deals with the creation of the corpora used by the Eurolect Observatory project for verifying the existence of specific legal language varieties in EU law. In order to highlight the existing linguistic differences between the legal directives issued by the European institutions and national implementing measures in EU Member States, two different types of corpora (Barbera et al., 2007, p. 70) – named corpus A and corpus B respectively – were set up in compliance with the standard methodology (Reppen, 2010).

A common protocol was conceived in accordance with the *Eurolect Observatory Research Project* goals (see Mori, this volume). By the effect of this agreement, a total of 22 corpora (11 for each of the A and B sections, Kenning, 2010) were collected by UNINT, the project coordinator's institution. The overall work was distributed as follows: 16 corpora were produced by UNINT for Dutch, French, English, German, Italian, Maltese, Polish and Spanish; the University of Tampere was responsible for the Finnish versions; Ventspils University College prepared the Latvian ones and, eventually, the University of Corfu was in charge of the Greek corpora. As the project leader and project manager, UNINT is the owner of the

corpora, which were produced by the author of this paper in accordance with the project objectives and technical design specifications (see Mori, this volume). By contrast, the development of corpus search engines (see section 2.5) followed an autonomous design. It is the result of a concurrent project that was started on personal initiative in order to address the needs of the user to access and extract information from corpora in an easy and straightforward way.

Owing to the multiplicity of the orthographic systems currently in use across Europe, all texts were encoded in Unicode UTF-8 (Gillam, 2003). As regards the file format to download, plain text or HTML represented the best choice. Converting a binary file is, in general, a complex process that can produce unwanted artifacts or layout corruptions. Tables and multiple-column texts are particularly affected by this issue; therefore the choice of processing proprietary formats was only made if no other solution was available.

As a final note, it is relevant to stress that in compliance with project specifications, both corpora and search tools were designed for local and personal use only. Therefore, no web-based corpora management facilities were implemented.

2 Corpus collection

2.1 Corpus A

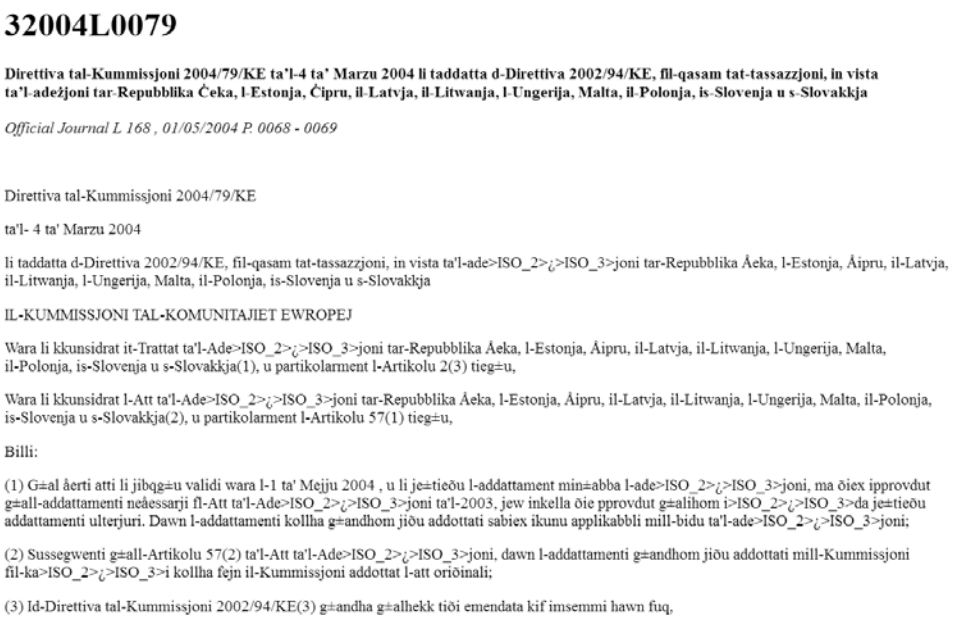
All versions of corpus A are collections of European Union Directives which were published in a time span of ten years, from 1999 to 2008. These norms were downloaded as text files in machine-readable format from the official EU law web portal Eur-Lex,[1] which is maintained by the Publications Office of the European Union. Because EU Directives indicate the policies that different EU Member States should put into force, their number may vary according to the subject dealt with. Probably because of this, the Maltese and Polish versions of corpus A contain only 656 and 658 texts respectively, while the other six languages count a total of 660 directives. More in particular, Directives 70/2005 and 29/2006 were not available in Maltese, while Directives 53/2001 and 97/2001 were unavailable in both Maltese and Polish.

Apart from the missing directives, the following issues were encountered:

- Directive 73/2004 - image PDF format (for all languages except French, Maltese and Polish).
- Directive 75/2004 - incomplete PDF file in German.
- Directive 79/2004 - bad text characters in Maltese (Figure 1).
- Directive 121/2006 - PDF format only.

[1] http://eur-lex.europa.eu/homepage.html

- Directive 101/2008 - marked as 101/2009 in Dutch (Figure 2).
- Bracketed directive identifiers were found in the titles of:
  Directives 87 and 137/2006.
  Directive 61/1999 - Polish version only.



32004L0079

Direttiva tal-Kummissjoni 2004/79/KE ta'l-4 ta' Marzu 2004 li taddatta d-Direttiva 2002/94/KE, fil-qasam tat-tassazzjoni, in vista ta'l-adeżjoni tar-Repubblika Ċeka, l-Estonja, Ċipru, il-Latvja, il-Litwanja, l-Ungerija, Malta, il-Polonja, is-Slovenja u s-Slovakkja

Official Journal L 168 , 01/05/2004 P. 0068 - 0069

Direttiva tal-Kummissjoni 2004/79/KE

ta'l- 4 ta' Marzu 2004

li taddatta d-Direttiva 2002/94/KE, fil-qasam tat-tassazzjoni, in vista ta'l-ade>ISO_2>¿>ISO_3>joni tar-Repubblika Åeka, l-Estonja, Åipru, il-Latvja, il-Litwanja, l-Ungerija, Malta, il-Polonja, is-Slovenja u s-Slovakkja

IL-KUMMISSJONI TAL-KOMUNITAJIET EWROPEJ

Wara li kkunsidrat it-Trattat ta'l-Ade>ISO_2>¿>ISO_3>joni tar-Repubblika Åeka, l-Estonja, Åipru, il-Latvja, il-Litwanja, l-Ungerija, Malta, il-Polonja, is-Slovenja u s-Slovakkja(1), u partikolarment l-Artikolu 2(3) tieg±u,

Wara li kkunsidrat l-Att ta'l-Ade>ISO_2>¿>ISO_3>joni tar-Repubblika Åeka, l-Estonja, Åipru, il-Latvja, il-Litwanja, l-Ungerija, Malta, il-Polonja, is-Slovenja u s-Slovakkja(2), u partikolarment l-Artikolu 57(1) tieg±u,

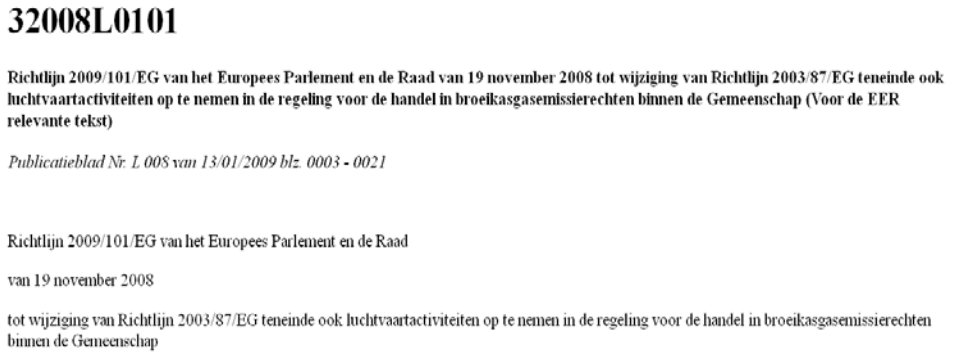
Billi:

(1) G±al åerti atti li jibqg±u validi wara l-1 ta' Mejju 2004 , u li je±tieðu l-addattament min±abba l-ade>ISO_2>¿>ISO_3>joni, ma ðiex ipprovdut g±all-addattamenti neåessarji fl-Att ta'l-Ade>ISO_2>¿>ISO_3>joni ta'l-2003, jew inkella ðie pprovdut g±alihom i>ISO_2>¿>ISO_3>da je±tieðu addattamenti ulterjuri. Dawn l-addattamenti kollha g±andhom jiðu addottati sabiex ikunu applikabbli mill-bidu ta'l-ade>ISO_2>¿>ISO_3>joni;

(2) Sussegwenti g±all-Artikolu 57(2) ta'l-Att ta'l-Ade>ISO_2>¿>ISO_3>joni, dawn l-addattamenti g±andhom jiðu addottati mill-Kummissjoni fil-ka>ISO_2>¿>ISO_3>i kollha fejn il-Kummissjoni addottat l-att oriðinali;

(3) Id-Direttiva tal-Kummissjoni 2002/94/KE(3) g±andha g±alhekk tiði emendata kif imsemmi hawn fuq,

Figure 1 – Directive 2004/79 displays unreadable characters in the Maltese version



32008L0101

Richtlijn 2009/101/EG van het Europees Parlement en de Raad van 19 november 2008 tot wijziging van Richtlijn 2003/87/EG teneinde ook luchtvaartactiviteiten op te nemen in de regeling voor de handel in broeikasgasemissierechten binnen de Gemeenschap (Voor de EER relevante tekst)

Publicatieblad Nr. L 008 van 13/01/2009 blz. 0003 - 0021

Richtlijn 2009/101/EG van het Europees Parlement en de Raad

van 19 november 2008

tot wijziging van Richtlijn 2003/87/EG teneinde ook luchtvaartactiviteiten op te nemen in de regeling voor de handel in broeikasgasemissierechten binnen de Gemeenschap

Figure 2 – Directive 2008/101 wrongly indicated as 2009/101 in the Dutch version

Despite the above-listed issues, corpus A proved to be the easiest data collection to build. After downloading the whole set of selected directives in the 8 languages UNINT was committed to dealing with, each single HTML file was automatically processed by a script in order to remove the markup tags, enrich the text with both header and structural information and, eventually, to save the result in a UTF-8 format file. During this task, the non-standard usage of brackets around the directive identifier was also properly managed.

As regards technical aspects, the script was developed by taking advantage of the GNU version of a text-oriented language[2] named AWK (Robbins, 2015). This interpreter, which is part of any Unix-like Operating System, operates by recursively applying the batch of statements defined in its program body to any text line the file to process consists of. The choice to adopt the AWK scripting language was made after evaluating its robustness, easiness and proven ability to manage a wide range of text mining and natural language processing tasks (Schmitt, Christianson & Gupta, 2007, pp. 221-258) for corpus linguistics (Burnage & Dunlop, 1992).

The first action accomplished by the processing script consists of adding a block header to all the texts of corpus A. Although it is not part of the text structure, the header covers the fundamental function of providing the user with clear metadata descriptions of the EU norm (Weisser, 2016, pp. 33-37). As such, the header is divided into two lines. The first line contains all the subjects the single EU directive deals with. This information is expressed in accordance with the official classification directory provided by Eur-Lex in all the languages in which the directive is published. Where multiple directory information is covered, it was convenient to divide the different couples of numbers and alphabetic values with a pipe " | " character, as shown in example 1. The second line of the header structure, instead, contains information about which EU Official Journal issue the original text can be found in (example 2). Like most laws, EU directives are also structurally divided into different functional areas. In particular, three main sections can be defined after the title. These are: "Preamble", "Disposition" and "Annex". As a consequence, in order to digitally represent such a partitioning, an XML-like[3] markup was provided. Example 3 shows the set of markup tags that were added to clearly define the textual organization of all the corpus directives involved in this project.

(1) <meta name="DC.subject" content="13.20.60.00 Industrial policy and internal market / Industrial policy: sectoral operations / Information technology, telecommunications and data-processing | 13.30.99.00 Industrial policy and internal market / Internal market: approximation of laws / Other sectors for approximation of laws">

(2) <meta name="DC.source" content="Official Journal L 066 , 13/03/1999 P. 0016 - 0023">

[2] http://savannah.gnu.org/projects/gawk/

[3] The mentioned tags are not to be considered as part of a real XML structure because no DTD was provided.

(3) <title>…</title>
<preamble>…</preamble>
<disposition>…</disposition>
<annex>…</annex>

As mentioned above, the entire cleaning and markup process was automatically managed by a script. However, for the program to work properly, the different HTML standards that the Eur-Lex Publication Office had adopted to produce the digital versions of the EU directives had to be taken into account. After completing a formal check of all files, we found that most of them were edited in HTML 4 (Figure 3), while some used the most recent XHTML format (Figure 4). Since the said standards produce different file structures, the need to implement two processing strategies inside the same script was mandatory. Although an easy way would have been to automatically remove the HTML tags from all directives by means of some free software, this solution could not be adopted for non-trivial reasons. Leaving aside all the operational aspects of such a tool, this process would produce a text file in which no difference between real content and other functional parts (like the file header or the document type declaration) could be identified. So, to correctly manage all the different types of information the file provided, the first element that the script had to check to correctly select the right algorithm to run was the HTML declaration. In case the “html xmlns:html” string was found, the program would have to process an XHTML file. Otherwise, commands for dealing with standard HTML 4 would be run.

Figure 3 – Directive published in HTML 4 standard

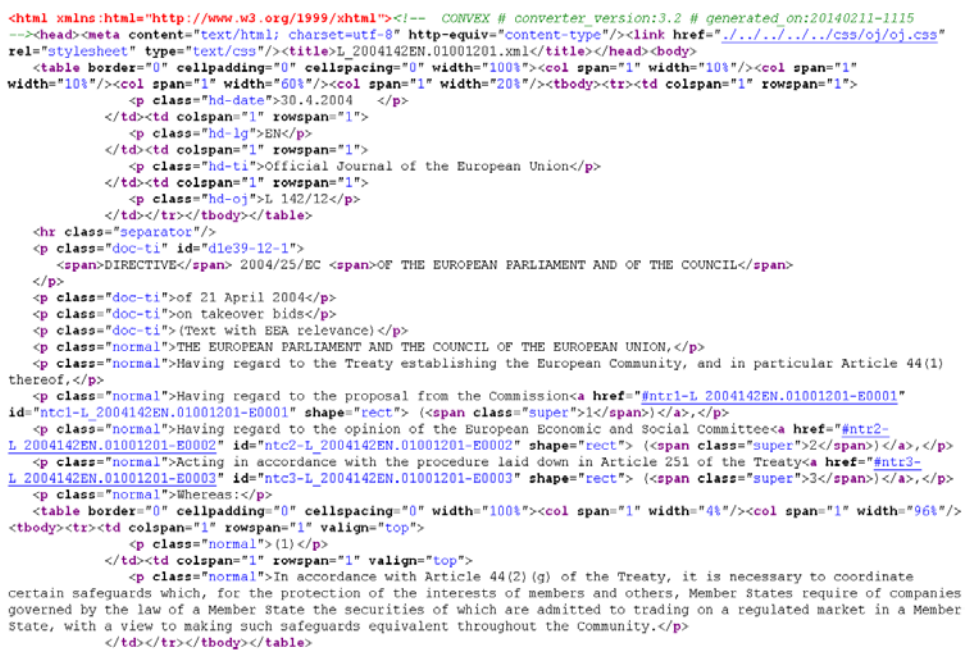


Figure 4 – Directive published in XHTML standard

Despite HTML file headers declaring that all the directives adopt the UTF-8 character set, in many cases the old HTML coded character set[4] could be found for representing accented letters and non-alphanumeric elements like mathematical symbols, apostrophes, quotation marks, and dashes (see Figures 5 and 6). Consequently, since for uniformity and usability reasons all our corpora had to include plain UTF-8 characters only, a conversion list was implemented.[5]

After this step, the script focuses on identifying the text macrostructure by taking into account both HTML code and text patterns like the usage of capital letters. Although EU directives are generally made up of three functional areas, in some cases the third component, the Annex, is not provided. Therefore, a control system was set up to prevent possible markup mistakes that could otherwise have been produced by the script. After removing the original markup tags, new structural information was added and specific procedures were run to correctly represent the original text layout. Because of the different HTML standards adopted, translating complex structures like tables and formulas from web format to plain text represented a difficult task that required specific processing.

Article premier 1. La présente directive établit les normes minimales relatives à la protection des poules pondeuses. 2. La présente directive ne s'applique pas: - aux établissements de moins de 350 poules pondeuses, - aux établissements d'élevage de poules pondeuses reproductrices. Ces établissements restent toutefois soumis aux exigences pertinentes de la directive 98/58/CE. Article 2 1. Les définitions figurant à l'article 2 de la directive 98/58/CE sont applicables pour autant que de besoin. 2. En outre, aux fins de la présente directive, on entend par: a) "poules pondeuses": des poules de l'espèce Gallus gallus ayant atteint la maturité de ponte et élevées pour la production d'oeufs non destinés à la couvaison; b) "nid": un espace séparé, dont les composants au sol excluent toute utilisation de treillis métalliques pouvant entrer en contact avec les volailles, prévu pour la ponte d'une poule ou d'un groupe de poules (nid collectif); c) "litière": tout matériel friable permettant aux poules de satisfaire leurs besoins éthologiques; d) "surface utilisable": une surface large d'au moins 30 centimètres, inclinée au maximum à 14 %, surmontée d'un espace libre haut d'au moins 45 centimètres. Les surfaces du nid ne font pas partie de la surface utilisable. Article 3 Selon le ou les système(s) retenu(s) par les États membres, ceux-ci veillent à ce que, outre les dispositions pertinentes prévues par la directive 98/58/CE et par l'annexe de la présente directive, les propriétaires ou détenteurs de poules pondeuses appliquent les exigences spécifiques à chacun des systèmes visés ci-dessous, à savoir: a)

Figure 5 – part of Directive 1999/74 - French version

[4] https://www.w3.org/MarkUp/html-spec/html-spec_13.html

[5] As an example, the HTML string "ĳ" codifies as a unique symbol the Dutch digraph "ij".

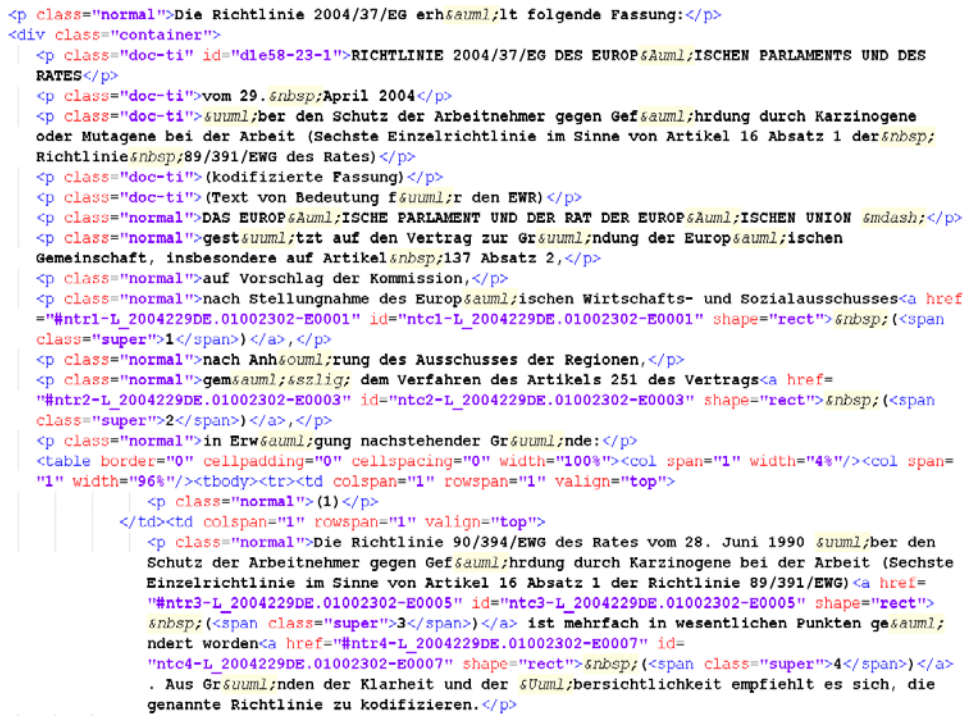


Figure 6 – part of Directive 2004/37 - German version

2.2 Corpus B

Corpus B is a collection of national implementing measures for EU directives. It complies with representativeness, parallelization and balance requirements because the file selection criteria are derived from those adopted for corpus A. However, in contrast to the EU directives, Member State laws are not hosted in a unique web repository. This made the corpus B file collection process more difficult to accomplish. Since the publication of the digital version of the domestic norms is directly managed by single government entities,[6] retrieving a definite set of digital resources from a specific website is not a simple selection-and-download activity. To perform this task, we had to master a number of practical aspects like the way in which the different web portals are organized, the possible formats in which the digitized texts were available for download, and the language each

[6] Domestic norms can be downloaded from the following sites:
England: http://www.legislation.gov.uk/
France: https://www.legifrance.gouv.fr/
Germany: https://www.bgbl.de/
Italy: http://www.gazzettaufficiale.it/
Malta: http://www.justiceservices.gov.mt/
Netherlands: https://www.overheid.nl/
Poland: http://www.dziennikustaw.gov.pl/
Spain: https://www.boe.es/

website adopts to communicate with users. This can be the most challenging issue because, in general, only the national language is used to take advantage of all the services the portal offers. A parallel English version is rarely provided, or it is limited to few general information pages (see Figure 7).

Figure 7 – Government of the Netherlands web portal - English version

In contrast to the relatively stable number of directives each EU member State has to enact, the amount of corresponding implementing measures is totally unpredictable because of a number of reasons hereafter discussed.

All the information about which Member State norms are designed to enact a specific directive is provided by a particular section inside the Eur-Lex portal. However, direct links to the domestic laws were not available at the time that data collection was carried out; moreover, in many cases law number referrals were incorrectly reported or missing.[7] In addition, the amount of usable norms was reduced by file format issues. In particular, our automatic procedures were not designed to directly process image PDF file content. To properly convert such data formats into machine-readable texts, the use of language-specific character recognition software (OCR) is mandatory. Despite their quality, all image converters suffer from technological limits that require users to perform time consuming post-editing to differing extents. Both the number of languages involved and the need to manually verify and correct all the output texts led to considering the OCR-based approach as an unfeasible solution because it would have caused serious project delays.

Different but more relevant problems affect Maltese. Probably because of the recent adoption of an independent alphabetic system, many

[7] Paradoxically, the Polish version of Directive 2001/53 was not available in Eur-Lex, but its national implementing measures were regularly reported for Poland.

laws are still available in English only (e.g. 112/2008). Yet, the real problem is tied to the way the official documents were produced. All Maltese norms can be downloaded in PDF format only. Although this would not present a problem if one limited oneself to reading their content, serious issues arise when trying to convert such files into another format like plain text. Because of the incomplete incorporation of the whole Maltese font set, a great number of Maltese implementing measures are unusable after their conversion (e.g. 362/2005 - see Figures 8 and 9). Actual data show that starting from an overall number of 445 original PDF files, only 139 of them - around 31% - could be converted into text and effectively used for research purposes. It is relevant to put in evidence that these problems were totally unpredictable and appeared only when the file processing activity was already started. This was caused by the fact that there is no way to know the specific PDF content format before opening or processing any single file. Because the only solution would require thorough manual revision to correct all the badly converted Maltese characters, no workaround could be implemented or planned in order to strictly comply with the project schedule.

A.L. 362 ta' l-2005

ATT DWAR IS-SERVIZZI VETERINARJI
(ATT NRU. XXIII TA' L-2001)

Regoli ta' l-2005 dwar it-Tqegħid fiċ-Ċirkolazzjoni ta' Ikel Kompost

BIS-SAHHA tas-setgha moghtija bl-artikolu 26 ta' l-Att ta' l-2001 dwar is-Servizzi Veterinarji, il-Ministru ta' l-Affarijiet Rurali u l-Ambjent għamel dawn ir-regolamenti li ġejjin:-

1. (1) It-titolu ta'dawn ir-regoli huwa Regoli ta' l-2005 dwar it-Tqeghid fiċ-Cirkolazzjoni ta' Ikel Kompost. Titolu u skop.

(2) Dawn ir-regoli ghandhom japplikaw ghal ikel kompost li jkun ghall-bejgh f'Malta.

(3) L-iskop ta' dawn ir-regoli hu l-implimentazzjoni tar-regoli misjuba taht id-Direttiva tal-Kunsill ta' l-Unjoni Ewropea 79/373/KE dwar it-tqeghid fiċ-ċirkolazzjoni ta' ikel kompost, hawnhekk aktar 'il quddiem imsejha "id-Direttiva".

(4) Dawn ir-regoli ghandhom japplikaw minghajr preġudizzju ghal dispożizzjonijiet dwar: –

(a) materjal ta'ikel;

(b) addittivi użati fl-ikel;

Figure 8 – Law 362/2005 in Maltese: PDF format

Suppliment tal-Gazzetta tal-Gvern ta' Malta, Nru. 17,836, 4 ta' Novembru, 2005 Taqsima B

A.L. 362 ta' l-2005
ATT DWAR IS-SERVIZZI VETERINARJI
(ATT NRU. XXIII TA' L-2001)

Regoli ta' l-2005 dwar it-Tqeg[id fiç-?irkolazzjoni ta'
Ikel Kompost

BIS-SA{{A tas-setg[a mog[tija bl-artikolu 26 ta' l-Att ta' l-2001 dwar is-Servizzi Veterinarji, il-Ministru ta' l-Affarijiet Rurali u l-Ambjent g[amel dawn ir-regolamenti li ;ejjin>
1. (1) It-titolu ta'dawn ir-regoli huwa Regoli ta' l-2005 dwar Titolu u skop. it-Tqeg[id fiç-Cirkolazzjoni ta' Ikel Kompost.
(2)
Dawn ir-regoli g[andhom japplikaw g[al ikel kompost li jkun g[all-bejg[f'Malta.

(3)
L-iskop ta' dawn ir-regoli hu l-implimentazzjoni tarregoli misjuba ta[t id-Direttiva tal-Kunsill ta' l-Unjoni Ewropea 79#373#KE dwar it-tqeg[id fiç-çirkolazzjoni ta' ikel kompost, hawnhekk aktar 'il quddiem imsej[a "id-Direttiva".

(4)
Dawn ir-regoli g[andhom japplikaw ming[ajr pre;udizzju g[al dispo]izzjonijiet dwar> –

(a)
materjal ta'ikel<

(b)
addittivi u]ati fl-ikel<

Figure 9 – Law 362/2005 in Maltese: text format

Another aspect we had to take into account regards the way the EU directives are implemented by national measures. Any State defines both the hierarchical level and the scope of each national law, so a clear distinction between the different legislation measures had to be drawn before downloading any file. According to our research project scope (see Mori, Introduction, this volume) only the laws that were directly promulgated by the national Parliament or were delegated to the Council of Ministers were taken into account as first choice. Consequently, both local norms and ministerial regulations were not included, unless specified by authors of chapters devoted to language cases for specific reasons. A typical example of this distinction is the case of Germany, where all the norms edited for the different Länder were left aside. The United Kingdom was affected by a similar problem. Only laws of national relevance for England were taken into account, while all the norms for Scotland, Northern Ireland and Gibraltar were not considered.

Nation	Number of Laws	Norm
United Kingdom	674	Acts of Parliament
Netherlands	504	Wet, Besluit
Germany	463	Bundesgesetze, Verordnung
Spain	438	Ley, Real Decreto, Real Decreto Ley
Italy	275	Legge, Decreto Legge
France	129	Loi, Ordonnance
Poland	482	Ustawa
Malta	139	Att tal-Parlament

Figure 10 – The primary norms in Corpus B

Since national implementing measures are independently managed by each single Member State, the relation between EU directives and National norms is not uniform across Europe. In general the EU directives are put into force over a time span of 3 to 4 years; this is the reason why in this project we have not used any directive published after 2008. In Poland, for example, Ustawa 2006-171-1225 was designed to enact a total of 108 directives from 1999 to 2008. On the other hand, a single directive can be enacted by a very specific part of a broader and more general law. For example, Directive 1999/062 is managed in France by article 87 of the Loi 98-546.

As regards non-textual information, corpus B adopts the same structural markup as corpus A. The corpus B header lines are also very similar to corpus A. However, their structure is more complex because a link between the two corpora had to be set to enable cross-corpus text search.

The first header line contains the list of subjects the National law enacts (see example 4). A single local law may put into force one or more directives. This information is stored in the second header line (see example 5). Finally, the third line contains the year the National norm was promulgated (see example 6).

(4) <meta name="DC.subject" content="13.30.10.00 Industrial policy and internal market / Internal market: approximation of laws / Motor vehicles | 07.20.40.10 Transport policy / Inland transport / Structural harmonisation / Technical and safety conditions">

(5) <meta name="DC.id" content="1999/7/EC , 1999/14/EC , 1999/15/EC , 1999/16/EC , 1999/17/EC , 1999/18/EC">

(6) <meta name="year" content="1999">

3 Corpus Search tools

3.1 Overview of the SearchIt tools

In the following we will describe from a technical point of view the three search engine modules that have been developed to help scholars effectively search and extract information from the *Multilingual Eurolect Observatory* corpora. Particular attention will be paid to pointing out the differences of the two engines in terms of search capabilities.

Since corpus A and corpus B are characterized by a different block of header structure, using two independent search engines is required. In addition, for the user to be able to run a query on both corpora by setting common selection criteria, designing a third search tool was required. Acting as a database query engine, this third program joins both corpora together by matching their metadata information. After doing this, all the text samples that satisfy the user-defined search pattern are extracted from corpus A and B. This task can be easily accomplished by reading the information provided by both the structural markup and header lines.

Like all the scripts that have been designed to build the different project corpora, the search engines described in this section were also developed by using AWK. Because SearchIt tools establish a question-based interaction with users, entire parts of program code are run multiple times. So, to avoid useless repetition, each script has been divided into several functional blocks that are called up as needed. As a consequence of this, no real program body has been implemented. Although unusual, the script structure is made up of a very simple BEGIN rule – it is sufficient to call the main program function as the program starts – and four user-defined functions.

Like other programs based on scripting languages, SearchIt is not an executable stand-alone tool. It can be run in a Unix-like shell environment, MS-DOS prompt or Windows command line by entering either the whole line required by the AWK interpreter or just the name of the search tool. This second option is granted by the so-called *shebang* line, which is placed at the beginning of the script.

3.2 Main functions of the SearchIt tools

The basic function of the BEGIN rule consists in managing the dynamic interaction with users in order to collect all the data that are required to correctly operate the text search-and-retrieve operations. As mentioned before, a number of questions regarding different parameters that the search tool has to manage are put forth to the user, asking for the following information (Figure 11):

1. Search result output file name. If no string is provided, the search results will be printed on the screen only.
2. Specific text section from which to retrieve data: preamble, disposition, annex. “All sections” is selected by default if no choice is made.
3. Corpus subject. Both EU directory classification codes and alphabetic strings can be entered. Optional. No subject-based filter will be used if “enter” is pressed with no input data.
4. EU directive publication time. The parameter may be a single year, a specific range, or the entire time span from 1999 to 2008. An error message is returned if the time frame is incorrectly entered. If no specific choice is made, the whole time span is selected by default.
5. Keyword, phrase or directory of EU legislation chapter number to search within the directive title. Optional. All texts will be taken into account if no input is provided.
6. Keyword to search in the selected texts.

After dealing with the above selections, the program will ask if case sensitivity and exact or partial match will be used during search. Next, the script will check whether the corpus A text file can be regularly accessed. In case of positive response, a number of variables are set in accordance with the selected criteria. Leaving aside search values referred to a particular text section and its content – actually this choice can operate without other information needed – the following seven combinations acting as a filter on both the directive header and title are considered:[8]

- publishing time
- subject
- title
- time and subject
- title and subject
- time and title
- time, title and subject

[8] The total number of possible combinations can be calculated using the formula $c=2^{p-1}$ where "c" indicates combinations and "p" the combining parameters. This figure has been reduced by 1 because the combination comprising no parameter selection, although possible, has not been considered.

```
$ ./searchit_a.awk
Output file name? (press 'Enter' to skip)
text.txt
Corpus section? (1 = preamble, 2 = disposition, 3 = annex - 'Enter' = all)
2
Corpus subject? ('Enter' = all)
food
Year range? (e.g. '2001-2003' , '1999' , 'Enter' = all)

Word or phrase to search in title? ('Enter' = skip)
food
Word to search in corpus? ('Enter' = skip)
food
Case sensitive? (y/n - 'Enter' = no)

Exact word match? (y/n - 'Enter' = no)
```

Figure 11 – Interactive selection of search criteria for corpus A

Data are retrieved if at least one search parameter is provided by the user. Therefore, if only default values are selected, an error is reported and the user is asked to choose between starting the search again or closing the program (Figure 12). If only the text section is selected, the tool will select and retrieve that specific part throughout the whole corpus. The number of texts retrieved is reported too (Figure 13).

```
$ ./searchit_a.awk
Output file name? (press 'Enter' to skip)

Corpus section? (1 = preamble, 2 = disposition, 3 = annex - 'Enter' = all)

Corpus subject? ('Enter' = all)

Year range? (e.g. '2001-2003' , '1999' , 'Enter' = all)

Word or phrase to search in title? ('Enter' = skip)

Word to search in corpus? ('Enter' = skip)

Warning: no word, section, title, subject or time range selected.
Press 'Enter' to restart or 'q' to quit
```

Figure 12 – Invalid search criteria

```
<disposition>
HAS ADOPTED THIS DIRECTIVE:
Article 1
Appendix II to Annex II to Directive 2006/87/EC is amended as set out in the A
nnex to this Directive.
Article 2
Member States which have inland waterways as referred to in Article 1(1) of Di
rective 2006/87/EC shall bring into force the laws, regulations and administra
tive provisions necessary to comply with this Directive with effect from 30 De
cember 2008. They shall forthwith communicate to the Commission the text of th
ose provisions.
When Member States adopt those provisions, they shall contain a reference to t
his Directive or be accompanied by such a reference on the occasion of their o
fficial publication. Member States shall determine how such reference is to be
 made.
Article 3
This Directive shall enter into force on the day of its publication in the Off
icial Journal of the European Union.
Article 4
This Directive is addressed to the Member States which have inland waterways a
s referred to in Article 1(1) of Directive 2006/87/EC.
Done at Brussels, 19 December 2008.
For the Commission
Antonio Tajani
Vice-President
[1] OJ L 389, 30.12.2006, p. 1.
[2] OJ L 373, 31.12.1991, p. 29.
</disposition>

660 document(s) found

Search again? (y/n - 'Enter' = y)
```

Figure 13 – No keyword, section-based search result

As mentioned before, all search criteria are stored in variables that will be used later to perform text selection activities. According to the program architecture, different logical blocks evaluate the user's choices and pass all the relevant data to a specific function named "analyse". This part of the program uses the received information to locate the text area to scan across the whole corpus and, subsequently, to pinpoint the selected keyword. The search engine has been designed to output all the lines containing the searched string both to the screen and to a file. A single line will be printed on the screen, while the entire context will be sent to the user-defined output file. As for the output data format, it is relevant to state that besides the text, header metadata and markup tags are printed as well.

Particular attention has been paid to managing the title properly. Since within the corpora a single title may take up multiple lines, it has been delimited by structural markup tags. If such a simple solution is adequate to describe the text blocks at a corpus level, it might be misleading and create problems in case third-party corpus analysis tools are used. So, in order to keep the title in the output and, at the same time, reduce to a minimum the risk of data misinterpretation because of the corpus structure, the whole title has been converted into a one-line header element introduced by the tag "DC.title" (Figure 14, line 3). Whereas creating the output file is regularly managed by the "analyse" function, in order to extract and print to screen the text lines the keyword is found in, a dedicated piece of program is called. Particular ANSI escape sequences ("\033[7m" and "\033[0m") are adopted to highlight the keyword and enhance its on-screen readability (Figure 14). The retrieved string is then returned to the calling function and

used for calculating statistical figures reporting the total amount of keywords found and which directive they are in (Figure 15).

```
<meta name="DC.subject" content="13.30.14.00 Industrial policy and internal ma
rket / Internal market: approximation of laws / Foodstuffs">
<meta name="DC.source" content="Official Journal L 066 , 13/03/1999 P. 0016 -
0023">
<meta name="DC.title" content="DIRECTIVE 1999/2/EC OF THE EUROPEAN PARLIAMENT
AND OF THE COUNCIL of 22 February 1999 on the approximation of the laws of the
 Member States concerning foods and food ingredients treated with ionising rad
iation">
<disposition>
1. This Directive shall apply to the manufacture, marketing and importation of
 foods and food ingredients, hereinafter called `foodstuffs`, treated with ion
ising radiation.
(a) foodstuffs exposed to ionising radiation generated by measuring or inspect
ion devices, provided that the dose absorbed is not greater than 0,01 Gy for i
nspection devices which utilise neutrons and 0,5 Gy in other cases, at a maxim
um radiation energy level of 10 MeV in the case of X-rays, 14 MeV in the case
```

Figure 14 – Keyword-based search result

```
184 result(s) found in 17 document(s)
word: food. found in document: 2006/125/EC
word: foods. found in document: 2003/13/EC
word: "Foodstuffs found in document: 2003/52/EC
word: foods found in document: 2006/125/EC
word: Food found in document: 2003/114/EC
word: Food found in document: 2007/19/EC
word: foods found in document: 1999/21/EC
word: food found in document: 2006/125/EC
word: foodstuff; found in document: 2003/114/EC
word: food found in document: 2003/114/EC
word: food found in document: 2005/79/EC
word: "food found in document: 2002/46/EC
word: foodstuffs found in document: 2007/42/EC
word: foodstuffs found in document: 2004/14/EC
word: food. found in document: 2007/19/EC
word: Food found in document: 1999/2/EC
word: foodstuffs found in document: 2005/31/EC
```

Figure 15 – Search statistics report

At the end of the search activity - or if no result is found (Figure 16) - a function is called to check the user's intention to start another search session. The program closes if a negative answer is given, otherwise the main function is run. As regards the file containing the search results, it is important to note that all three search tools are designed to flush data from the computer memory to the filesystem and let the user view the search results only after the program is closed. If a user decides to go on and search again, output results will be appended to the previous ones, at the bottom of the file.

```
$ ./searchit_a.awk
Output file name? (press 'Enter' to skip)
text.txt
Corpus section? (1 = preamble, 2 = disposition, 3 = annex - 'Enter' = all)
2
Corpus subject? ('Enter' = all)
environment
Year range? (e.g. '2001-2003' , '1999' , 'Enter' = all)

Word or phrase to search in title? ('Enter' = skip)
vegetable
Word to search in corpus? ('Enter' = skip)
lea
Case sensitive? (y/n - 'Enter' = no)

Exact word match? (y/n - 'Enter' = no)


Sorry, word not found

Search again? (y/n - 'Enter' = y)
```

Figure 16 – No result found

The SearchIt tool for corpus B differs from the version for corpus A in one search parameter only. Actually, in both corpus A and B most structural features are identical. Nonetheless, corpus B text headers hold two different publication date information (see section 2.2 - examples 5 and 6). Data can be searched not only by looking at the domestic norm publication date, but also by taking into account the date of the EU directive it implements. Although, from an operative perspective, this feature only represents a different request the user should respond to, from a programming point of view introducing this new parameter has to be dealt with in a proper way. Similarly to the directive's subject and publication date, the national transposition measure implementation date acts as a filter at the header line level. Even though at first sight this would represent an increase in the non-textual search combinations from 7 to 15, in practice the said combinations remain unchanged. It can be so because in corpus B the information held by the norm title is not relevant; therefore it will not be considered a part of the search activity (Figure 17).

```
$ ./searchit_b.awk
Output file name? (press 'Enter' to skip)
text_b.txt
Corpus section? (1 = preamble, 2 = disposition, 3 = annex - 'Enter' = all)
2
Corpus subject? ('Enter' = all)

Year range of EU Directives? (e.g. '2001-2003' , '1999' , 'Enter' = all)

Year range of National laws? (e.g. '2001-2003' , '1999' , 'Enter' = all)

Word to search? ('Enter' = skip)
food
Case sensitive? (y/n - 'Enter' = no)

Exact word match? (y/n - 'Enter' = no)
```

Figure 17 – Interactive selection of search criteria for corpus B

Differently from the tools described so far, the third search engine has been designed to simultaneously retrieve data from both corpus A and B with a single command. It can do so by exploiting the directive date/number value, which is the only metadata element the two corpora have in common besides their structural similarities. This means that the directive date/number information from corpus A can be used to create a link between the two corpora and retrieve the related texts from corpus B. Actually, the search engine operates in one direction only; it starts from corpus A and moves to corpus B. This implies that the user's selection criteria are requested only once and are used to retrieve texts from both corpora without any need for further parameters. The system will only ask for the user intention to run the search in corpus B and for a filename where the output will be stored. Summing up, this last search tool is not just a program that sequentially runs the SearchIt A and B engines and combines the two outputs properly; it unifies the two search algorithms into a uniquely optimized text retrieval solution (Figures 18 and 19).

```
Output file name? (press 'Enter' to skip)

Corpus section? (1 = preamble, 2 = disposition, 3 = annex - 'Enter' = all)
2
Corpus subject? ('Enter' = all)
agri
Year range? (e.g. '2001-2003' , '1999' , 'Enter' = all)

Word or phrase to search in title? ('Enter' = skip)

Word to search in corpus? ('Enter' = skip)
veg
Case sensitive? (y/n - 'Enter' = no)

Exact word match? (y/n - 'Enter' = no)


<meta name="DC.subject" content="03.50.20.00 Agriculture / Approximation of la
ws and health measures / Plant health | 03.60.51.00 Agriculture / Products sub
ject to market organisation / Cereals">
<meta name="DC.source" content="Official Journal L 075 , 14/03/2006 P. 0007 -
0016">
<meta name="DC.title" content="Commission Directive 2006/30/EC of 13 March 200
6 amending the Annexes to Council Directives 86/362/EEC, 86/363/EEC and 90/642
/EEC as regards maximum residue levels for the benomyl group (Text with EEA re
levance)">
<disposition>
In Annex I to Directive 90/642/EEC in category "2. Vegetables, fresh or uncook
ed, frozen or dry, (iii) Fruiting vegetables, (a) Solanacea", the entry "okra"
 is added between the entries "aubergine" and "others".
</disposition>
```

Figure 18 – Search results from corpus A

```
Search the results in the Corpus of national measures? (y/n - 'Enter' = y)
y
Output file name? (press 'Enter' to skip)


<meta name="DC.subject" content="03.50.40.00 Agriculture / Approximation of la
ws and health measures / Seeds and seedlings">
<meta name="DC.id" content="1999/105/EC">
<meta name="year" content="2002">
<meta name="title" content="Statutory Instruments 2002 No. 3026 SEEDS The Fore
st Reproductive Material (Great Britain) Regulations 2002 Made 4th December 20
02 Laid before Parliament 9th December 2002 Coming into force 1st January 2003
">

<meta name="DC.subject" content="03.50.20.00 Agriculture / Approximation of la
ws and health measures / Plant health">
<meta name="DC.id" content="2000/29/EC">
<meta name="year" content="1993">
<meta name="title" content="Statutory Instruments 1993 No. 1320 PLANT HEALTH T
he Plant Health (Great Britain) Order 1993 Made 20th May 1993 Laid before Parl
iament 21st May 1993 Coming into force 1st June 1993"

<meta name="DC.subject" content="03.50.20.00 Agriculture / Approximation of la
ws and health measures / Plant health | 03.50.40.00 Agriculture / Approximatio
n of laws and health measures / Seeds and seedlings | 13.30.99.00 Industrial p
olicy and internal market / Internal market: approximation of laws / Other sec
tors for approximation of laws">
<meta name="DC.id" content="2008/90/EC">
<meta name="year" content="2010">
<meta name="title" content="Statutory Instruments 2010 No. 2079 Plant Health S
eeds The Marketing of Fruit Plant Material Regulations 2010 Made 11th August 2
010 Laid before Parliament 19th August 2010 Coming into force 17th September 2
010"

3 document(s) found
```

Figure 19 – Search results from corpus B

Obviously, this third search tool is designed for token-based inquiries. If the user intends to run a subject-based selection of directives or national transposition measures, the use of SearchIt dedicated to the single corpus A or corpus B is suggested.

References

Burnage, G., & Dunlop, D. (1992). Encoding the British National Corpus. In Jan Aarts, Pieter de Haan, & Nelleke Oostdijk (Eds.), *English language corpora: design, analysis and exploitation. Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen 1992* (pp. 79-95). Amsterdam: Rodopi.

Gillam, R. (2003). *Unicode demystified: a practical programmer's guide to the encoding standard*. Boston: Addison-Wesley.

Barbera, E., Corino, E., & Onesti, C. (2007). Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup. In Emanuele Barbera, Elisa Corino, & Cristina Onesti (Eds.), *Corpora e linguistica in rete* (pp. 25-88). Perugia: Guerra Edizioni.

Kenning, M.-M. (2010). What are parallel and comparable corpora and how can we use them? In Anne O'Keeffe & Michael McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 487-500). London: Routledge.

Lenci, A., Montemagni, S., & Pirrelli, V. (2016). *Testo e computer. Elementi di linguistica computazionale*. Roma: Carocci.

Reppen, R. (2010). Building a corpus: what are the key considerations? In Anne O'Keeffe & Michael McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 31-37). London: Routledge.

Robbins, A. (2015). *Effective AWK programming: Universal text processing and pattern matching*. Sebastopol, CA: O'Reilly Media.

Schmitt, L. M., Christianson, K., & Gupta, R. (2007). Linguistic computing with UNIX Tools. In Anne Kao & Steve R. Poteet (Eds.), *Natural language processing and text mining* (pp. 221-258). London: Springer.

Weisser, M. (2016). *Practical corpus linguistics: an introduction to corpus-based language analysis*. Hoboken, NJ: Wiley.