[*Translating and the Computer 27*, November 2005 [London: Aslib, 2005]

MyCatex
A language-independent term extractor

José Vega (jveqa@mv-xml.com)
Benoît Lamey (blamey@etg.be)

my-xML
(www.my-xml.com)

Luxembourg

Jacques Vergne
(Jacques.Vergne@info.unicaen.fr)

GREYC, Université de Caen
(http://www.qreyc.unicaen.fr/)

France

Outline

This paper presents my-xML's Candidate Term Extractor (MyCatex), a language independent term extractor that works without any language-specific resources. MyCatex is currently developed by my-xML, a Luxembourgish language-engineering company specialized in multilingual content management (http://www.my-xml.com/).

The MyCatex term extraction algorithm is based on the current researches of Jacques Vergne from the University of Caen in France. MyCatex can be used for term extraction, semi-automatic and automatic generation of multilingual thesauri and document tagging and classification.

The following chapters introduce:

√ the "General Architecture" of MyCatex and how the outputs can be exported towards third part applications

√ the "Extraction" chapter explains the broad outlines of the algorithms developed for extracting candidate terms from texts in different languages

√ the "Validation" process chapter lists the validation functionalities of MyCatex used by the linguists/terminologists/documentalists in order to validate or invalidate the extraction results suggested by the software

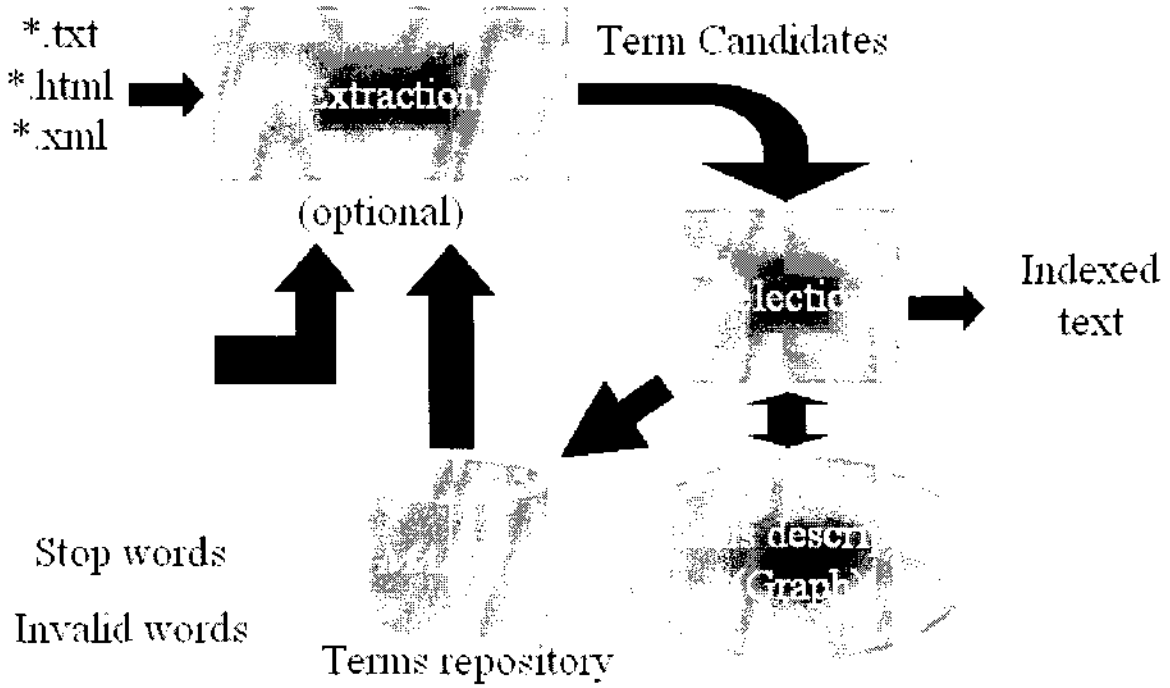
√ the "Terms Description" process chapter lists the various ways to add and visualise syntax and semantic information to the selected terms

√ the "Terms Repository" chapter describes the data format used to store and re-use the selected terms

√ the chapter "Corpus" explains how we proceed for functional and regression test automation.

General architecture

The software architecture is the following:



First the text is loaded in the software for extraction. The extraction engine doesn't require any language dependant information. However, some language-specific information can be added in order to improve the extraction results. For instance: stop words, invalid words, validated terms...

Then, the extraction engine generates a sorted list of candidate terms. The validation of these terms can be either automatic or semi-automatic. For semi-automatic validation, a set of tools is provided to help the user in the validation process.

The selected terms can be stored in a term repository or used to index the source text.

When saving terms in the Terms Repository, you can also add syntactical and semantic information. The Terms Description process allows the user to add and visualise this information in a graphical way showing relationships between terms.

The Terms Repository uses the XML format (respecting MARTIF standards for the representation of multilingual terminological data ISO-12620 DATA CATEGORIES). This allows the integration of these terms in third party applications or my-xML products, such as MyTerm[1] and MyTerm-Glossy[2] (further information about these products on www.my-xml.com).

[1] - Multilingual resources manager, cross-language search engine, multilingual content management system.

[2] - Computer aided translation tool for text and HTML documents. The user interface is the Word® editor.

MyCatex is multi-platform (coded in Java). Currently the interface is available in French and English and it can be localised easily in other languages since the menu of the interface is stored in XML .

Extraction

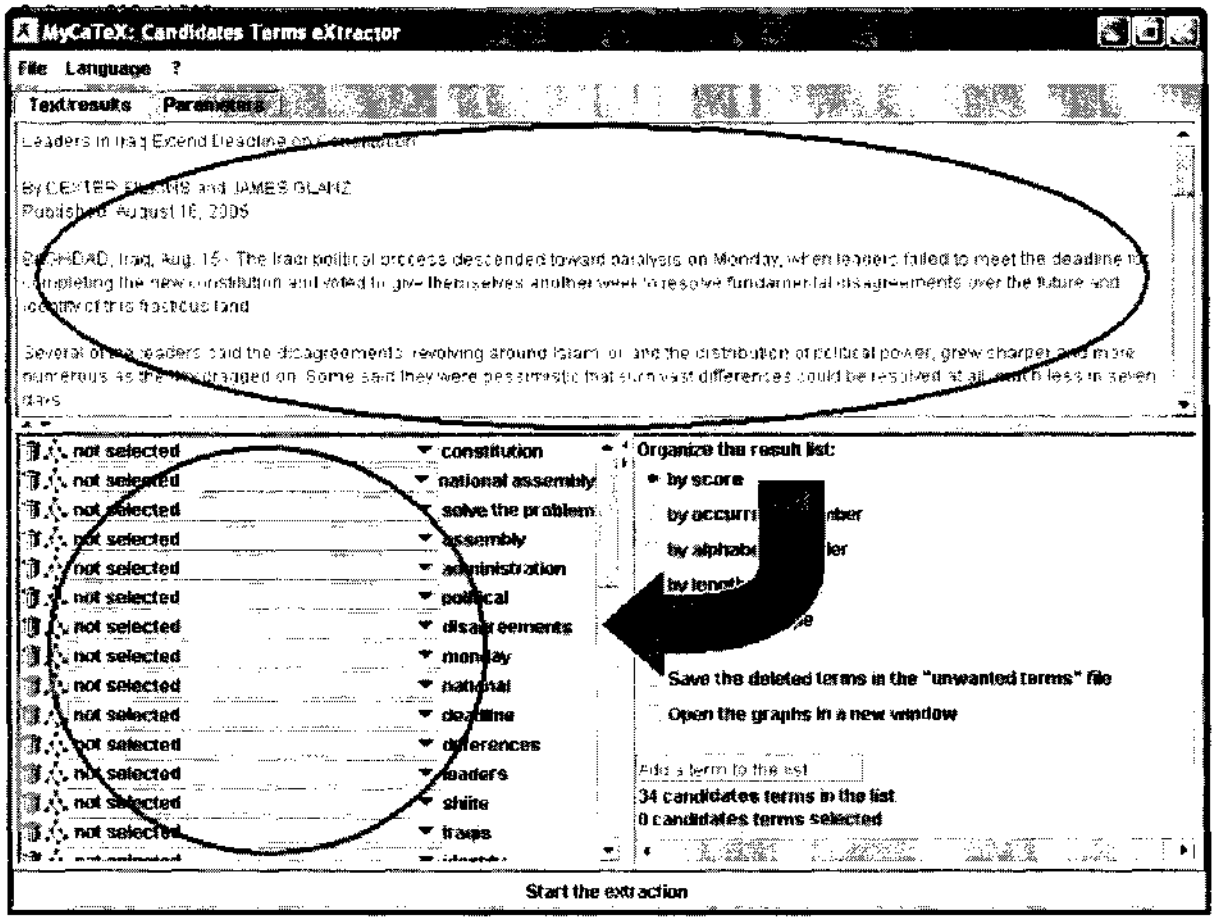
The development of the MyCatex extraction engine is the result of a cooperative work between my-xML and Prof. Jacques Vergne from the University de Caen (France). The extraction algorithm is based on the words sequences recurrence, and on the structure control of these sequences. This structure control is based on tagging the text to index with *a* two label "tagset": informative words or not. Informative words are defined as longer and less frequent than their neighbours. Very general linguistic properties, discovered by Zipf and by Saussure are thus exploited.

The extraction process can be divided into 4 steps:

1) Document tagging with 2 tags: informative words (I) and non informative words (n). We propose the following definition for an informative word: an informative word is longer and less frequent than the neighbouring words.
2) Generation of candidate terms based on the previous tagging: A candidate term doesn't include any punctuation mark. The candidate terms are those that follow the patterns I+, I+n+I+ or I+n+I+n+I+. The pattern definitions are based on regular expression syntax.
3) Terms grouping: the candidate terms that are enclosed in other candidates are grouped together, i.e. "Presidency of the European Union" will be grouped together with "Presidency" and "European Union".
4) Weighting of the candidate terms: the weight of each term candidate is the multiplication of the term's length by the number of occurrence of the term. This weighting is used in the validation process.

On the interface, the extraction is done as follow:

First a text is selected and appears in the upper part of the interface, and then the "extraction" button is pressed. The result is a list of candidate terms as shown in the screenshot hereafter.



Various parameters (optional) can be activated in the extraction process:

Extraction parameters

Selected patterns:

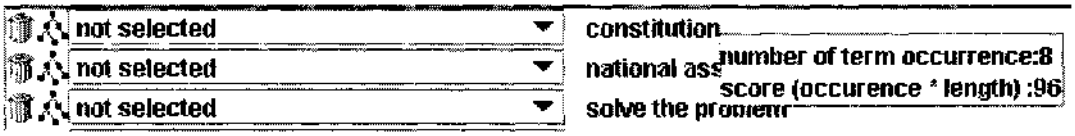
I+	☑
I+n+I+	☑
I+n+I+n+I+	☑
maximum size of the extracted terms (number of chars)	0
maximum size of the extracted terms (number of words)	0
minimum size of the extracted terms (number of chars)	0
minimum score for the extracted terms	0
Minimum occurrence number for a valid term candidate	2
maximum number of terms to extract	0
Loaded keyword file:	
empty words list	
list of words to exclude	
Unwanted terms file:	

Validation

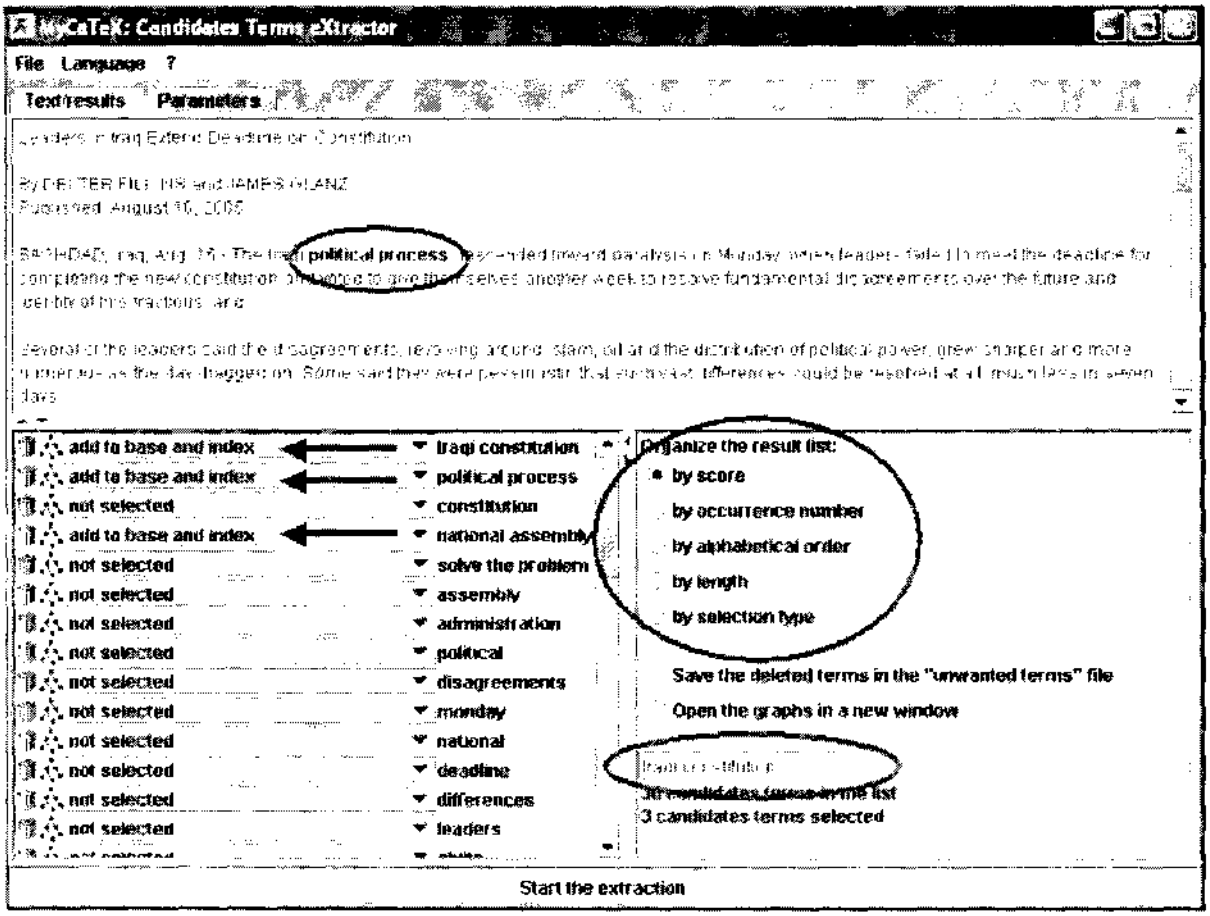
The validation process can be automatic by using the raw output of the extraction engine, or by adding some filters to the generated output (i.e. select the first 10 candidates) Some external linguistic resources (in this case language dependant) can be used to improve the results such as stop words lists or external terms repository (for automatic pre-selection of terms).

The validation can also be done manually using the same filters plus a human verification. This verification is done through a graphical interface. A term can be validated as index of the source text, as entry for the terms repository or both.

The user can view statistical information about the term (length, weight, number of occurrences).



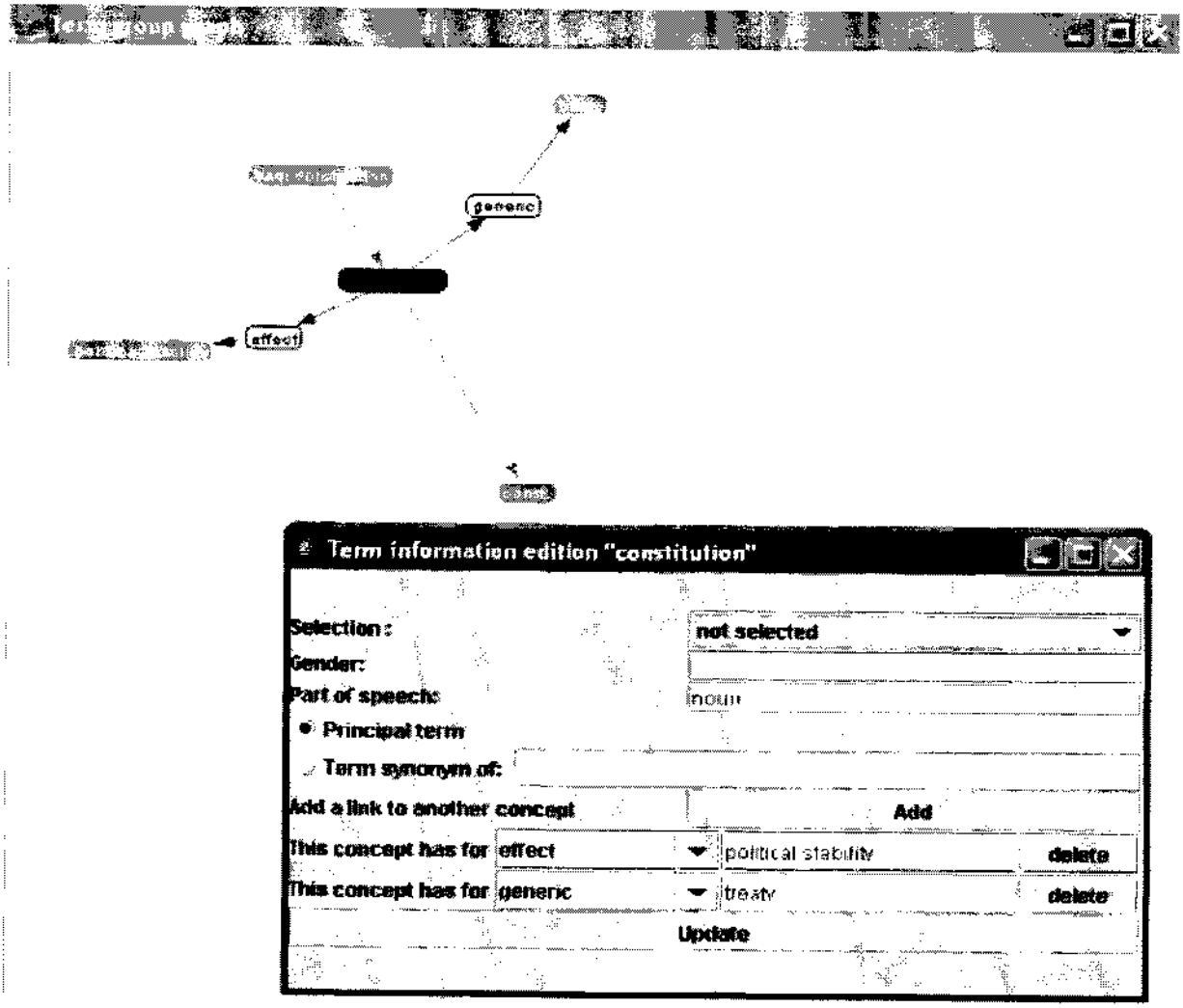
It is also possible to select/unselect/erase terms, sort the list of results according to various parameters, browse the text between the various occurrences of the candidate terms and select additional terms in the text.



Terms description

Terms description is used to add or visualise syntax information (part of speech or gender) and relations between terms (synonyms, association, opposite, cause, effect, generic, specific etc.)

This information is represented as a graph. The graph also shows the grouping of concatenated terms:



On this example we have "Iraqi constitution" which contains "constitution". "Constitution" is defined as a "noun" and has semantic links with "political stability" and "treaty".

Terms repository

The purpose of the term repository is to save valid terms which can be reused for further extraction. Terms are stored in XML respecting the MARTIF format. This allows the terms to be fed into a terminological database, a third party software or in my-xML applications. The saved terms can be organized into domains. The domains are defined by the context of the project.

The representation of a term in the database is as follow:

```
<termEntry id="2" type="conceptentry">
        <langSet lang="fr">
        <ntig>
        <termGrp id="2fr1">
                        <termnote type="termtype">main entry</termnote><termnote
                type="partOfSpeech">noun</termnote>
                        <term id="2fr1">constitution</term>
                        <date type="modification">2005-08-17T15:44:08</date>
                        <date type="input">2005-08-17T15:36:29</date><refObject
                type="subjectFieldSet">Politics</refObject><ptr type="updater" target="John" /><ptr
                type="originator" target="Ulla" /></termGrp>
                </ntig></langSet>
<ref target="political stability" type="effect" />
<ref target="treaty" type="generic" />
</termEntry>
```

Corpus and test

The corpus on which the results of MyCatex are currently tested is the "Acquis Communautaires" corpus (AC). It contains 8000 legal texts from the UE, in each of the 20 official languages.

The AC and other Community legislation texts are publicly available on the European Commission's web sites. The Language Technology team of the *Joint Research Centre* in Ispra, Italy, has attempted to identify the documents that are part of the AC, has downloaded and converted the texts into XML format. In further processing steps, the texts were cleaned of their footers and annexes and sentence-aligned. Instead of using a single pivot language, all possible language pair combinations were aligned individually. This is useful for the n-to-n relationship between aligned sentences, which often differs depending on the language pair involved.

Here is an extract of one document of the corpus:

[...]<P sid="4">COUNCIL DECISION OF 2 APRIL 1963 LAYING DOWN GENERAL PRINCIPLES FOR IMPLEMENTING A COMMON VOCATIONAL TRAINING POLICY </P>

<P sid="5">(63/266/EEC) </P>

<P sid="6">THE COUNCIL OF THE EUROPEAN ECONOMIC COMMUNITY , </P>

<P sid="7">HAVING REGARD TO THE TREATY ESTABLISHING THE EUROPEAN ECONOMIC COMMUNITY , AND IN PARTICULAR ARTICLE 128 THEREOF ; </P>

<P sid="8">HAVING REGARD TO THE PROPOSAL FROM THE COMMISSION ; </P><P sid="9">HAVING REGARD TO THE OPINION OF THE ECONOMIC AND SOCIAL COMMITTEE ; </P>

<P sid="10">HAVING REGARD TO THE OPINION OF THE EUROPEAN PARLIAMENT (1) ; </P>

<P sid="11">WHEREAS , IN ORDER TO FULFIL THE OBLIGATION IMPOSED ON THEM BY THE TREATY , TO ENSURE THE MAINTENANCE OF A HIGH LEVEL OF EMPLOYMENT THROUGH THEIR ECONOMIC POLICIES , MEMBER STATES SHOULD TAKE APPROPRIATE ACTION TO ADAPT THE SKILLS OF THEIR LABOUR FORCES TO CHANGES IN THE GENERAL ECONOMIC SITUATION AND TO CHANGES IN PRODUCTION TECHNOLOGY ; </P>

<P sid="12">WHEREAS AGAINST THE BACKGROUND OF THE RAPID IMPLEMENTATION OF THE COMMON MARKET AND IN CONJUNCTION WITH THE CO-ORDINATION OF REGIONAL POLICIES AND THE PROGRESSIVE ACHIEVEMENT OF A COMMON AGRICULTURAL POLICY , THE STRUCTURAL CHANGES WHICH ARE AT PRESENT TAKING PLACE IN CERTAIN ECONOMIC SECTORS RAISE URGENT PROBLEMS OF VOCATIONAL TRAINING AND RETRAINING ; </P>[.....]

On the next page we show part of the extraction results for this document and the French, Dutch, Polish and Greek version.

Beware, in the table hereafter the candidate terms are displayed according to the heuristics used by the extractor (term's length by the number of occurrence). It is not an aligned multilingual terminology.

EN	FR)	NL	PL	GR
common vocational training policy	politique commune de formation professionnelle	gemeenschappelijk beleid met betrekking tot de beroepsopleiding	zawodowego	καταρτίσεως
vocational training policy	formation	beroepsopleiding	ustanawiająca ogólne zasady realizowania wspólnej polityki	εφαρμογή κοινής πολιτικής επαγγελματικής καταρτίσεως
advanced training of teachers	formation professionnelle	lid-staten	miejsca kształcenia i miejsca kształcenia i miejsca pracy	πολιτικής επαγγελματικής καταρτίσεως
member states	perfectionnement du personnel enseignant	toepassing	współpracy z państwami członkowskimi	κοινής πολιτικής επαγγελματικής καταρτίσεως
training of young persons	domaine de la formation professionnelle	commissie	realizowania wspólnej polityki kształcenia zawodowego	κράτη μέλη
principles for implementing a common vocational	états membres	onderwijzend personeel en van leermeesters	zatrudnieni na stanowiskach niższych niż kierownicze	τομέα της επαγγελματικής καταρτίσεως
common vocational	communauté	gemeenschap	miejsca kształcenia i miejsca kształcenia i miejsca	επαγγελματικής καταρτίσεως
commission	Politique	gebied van de beroepsopleiding	polityki kształcenia zawodowego	εφαρμογή κοινής πολιτικής
principles for implementing a common	commission	algemene beginselen	życie wspólnej polityki kształcenia zawodowego	2ας απριλίου 1963 περί θεσπίσεως των γενικών αρχών
implementing a common vocational	personnel enseignant	bevoegde instellingen van de gemeenschap	polityki kształcenia zawodowego szczególną	επαγγελματικής
particular		vaststelling van de algemene beginselen	państwa członkowskie	κοινή πολιτική επαγγελματικής καταρτίσεως πρέπει
community			ustanawiająca ogólne zasady realizowania	δύνανται να ασκήσουν επαγγελματική δραστηριότητα
suitable training programmes			szczególności	γενικών αρχών για την εφαρμογή κοινής πολιτικής
effective common vocational training policy			nauczycieli i instruktorów	γενικές αρχές για την εφαρμογή κοινής πολιτικής
co-operation with the member states			stanowiskach niższych niż kierownicze	απριλίου 1963 περί θεσπίσεως των γενικών αρχών

Further development

An overview of the future developments for MyCatex:

- √ Algorithms and parameters improvements in order to obtain better results (reduce noise)
- √ use of MyTerm as a terms repository
- √ implementation of the Eurovoc indexing system as domain reference for the terms
- √ combine use of MyCatex and aligned corpora to generate multilingual aligned terminology