Polish Monolingual Task within Cultural Heritage in CLEF (CHiC) 2013. Wrocław Runs.

Adam Pawłowski[1]

[1] Institute of Information and Library Science, University of Wroclaw, Wroclaw, Poland
apawlow@uni.wroc.pl

Abstract.
The objective of the experiment described in the article was the evaluation of Polish digital resources searchable by the engine Europeana. Queries were created so as to represent proportional shares of common names, named entities and their combinations. They were manually enriched by students, experienced professionals and humanities-oriented educated persons. The system responses were then evaluated according to their Mean Average Precision. The average efficiency of information retrieval for Polish monolingual queries was weak: there were only 26.6% of highly relevant responses and as much as 73.5% of queries produced unsatisfactory results. MAP produced the best results for the automatic search (0,314), while enriched files contained less relevant results: for the expert users MAP value was 0,1795, for the educated users it was 0,1529 and for students 0,1279. The overall results proved that the IR process concerning Polish resources searchable by Europeana require significant improvement.

Keywords. CHiC Polish Task, ad-hoc IR, OKAPI

1 Introduction and objective

Europeana is an ambitious project intended to integrate the Old Continent's digitised heritage preserved in multiple national collections and to deliver information about its content in users' languages [3]. This aim is nonetheless difficult to achieve. Unlike the majority of global information services which were created from the beginning as monolingual, Europeana strives to overcome the problem of extensive multilingualism of the resources and of its potential users. Another obstacle to the effective information retrieval here is the existence of a variety of domain-, language- and country-specific traditions of creating meta-descriptions of documents and/or exhibits. This specificity is noticeable even if all the "local" descriptions are converted into a unified format, e.g. Dublin core. Finally Europeana's communication interface should respect different heuristic strategies of potential users in the European countries, related to their native language, educational profile and expectations.

Europeana covers resources in thirty languages. Items or objects qualified as "Polish" are an important and constantly growing part of this offer as they constitute at present the ninth richest subset of the entire collection. Regarding this issue from the user's perspective one should say that native speakers of Polish – thus potential Europeana audience – are the fifth population of Europe, if only languages using Latin script are considered [9].

The objective of the experiment described below was to evaluate the effectiveness of information retrieval conducted on Polish digital resources with the help of the engine Europeana. By "Polish" items we understand documents of any kind, prepared in any language and of any origin (not only Polish), offered by libraries, museums, documentation centres etc. located in Poland.

2 Some linguistic characteristics of Polish

Polish is not commonly used in information retrieval systems and some information about its lexis and morphology is necessary here. As there exists a number of grammars of this language here only the features relevant for IR activities will be highlighted [1, 5, 7, 12]. Polish is a flectional language belonging to the West-Slavic group of Indo-European family. It uses Latin alphabet with a series of specific characters: the nasal vowels *ą, ę,* the vowel *ó*, the semi-vowel *ł* and the consonants *ć, ń, ś, ź, ż*. These characters may pose a two-fold obstacle to the IR systems. Technically such systems are not prepared to process them fully (they are automatically replaced by the corresponding Latin characters or rejected). As far as system users are concerned native speakers of Romance or Germanic languages in their communicative strategies have a similar tendency: they reduce them as unknown to the recognised Latin characters without diacritics. It should be thus stressed that replacing these specific characters by the corresponding plain Latin forms should be avoided, as it distorts communication and may lead in some cases to the change of sense (e.g adj. *sądowy* means *of the court,* while *sadowy* means *of the orchard*).

Polish lexical resources have a large number of common Indo-European stems and lexemes. There are numerous words of Latin and Greek origin in the scientific terminology. A number of direct or indirect German borrowings entered Polish in the Middle Ages and in the nineteenth century. Polish borrowed also hundreds of words from Italian (sixteenth century) and from French (eighteen century). Although spelling of these words is usually adjusted to the phonology and morphology of the "target language", their etymology is rather an advantage to users in a multilingual environment. A separate issue is spelling used in old documents and partly in their descriptions. Systems are not prepared to lemmatize or disambiguate old orthography but users who intend to work on such documents are usually familiar with the problem.

What seems to be the main challenge for the IR systems handling resources in Polish is undoubtedly morphology. Polish nouns and adjectives are inflected in number, case and gender. Polish verbs are conjugated according to person and tense. Verb tenses are additionally modified by aspect (e.g. perfect, iterative etc.) which is a systemic feature of Polish, contrary to Romance and Germanic languages. All the parts of speech have also developed a rich derivational system. Surprisingly for the users of western languages even proper names are inflected in Polish. This concerns all the *nomina propria* of Slavic origin and most of the borrowed ones.

There is a visible tendency to use analytical forms in Polish which means that grammatical compound morphemes are more and more often replaced by prepositions. For instance the expression *send to the president* could be rendered either as *wysłać prezydentowi* or as *wysłać* **do** *prezydenta* but the second form using the preposition *do* (more analytical) becomes prevalent. In spite of this tendency information retrieval systems based on natural language processing should consider the aforementioned characteristics of Polish as a highly flexional language. While verbs in document descriptions are often avoided or reduced to basic forms (present tense, third person, no aspect) because they do not carry significant information, nominal elements require special attention. They should be recognized and disambiguated with regard to grammatical case and number. It should be reminded that the most frequent cases in Polish are genitive, accusative and dative. Since several nouns may produce complex semantic derivatives the system should also interpret them properly. E.g. from the city name *Lublin* on can derive the adjective *lubelski*, the nouns *lublinianin*, *lublinianka* (male or female inhabitant of Lublin), and the region name *lubelszczyzna*, all of them inflected by number and case. But disambiguation of search terms includes also resolution of polysemy or homonymy. In this example one should distinguish between the city name and the car make of delivery vans of the same name.

3 Experiment setup

3.1 Collection

The so called “Polish collection” is a part of CHiC 2012, and 2013 Multilingual collection. It consists of 1,093,705 documents contained in 1,094 files and constitutes the ninth richest collection of all the 30 languages handled by Europeana. The whole archive has a size of 119 MB and was made available by Europeana last year at http://ims.dei.unipd.it/data/chic/. According to CHiC 2012 evaluation [8], Polish collection included in 2012 975,818 text documents, 117,075 images, 582 videos and 230 sound documents (cf. Tab. 1). It was assumed that there had been no significant changes in data since 2012 until the 2013 experiment.

Table 1. Structure of CHiC 2013 Polish collection (based on the CHiC 2012 data)

Media type	documents	percentage of collection
text	975,818	89.221%
images	117,075	10.704%
videos	582	0.053%
sound	230	0.021%

As described in [8] Europaena content is provided with metadata describing digital representations of cultural heritage objects. In order to achieve this all the collection files use a special XML format combining different schemas [2]:

- Dublin Core (all tags starting with *dc:* prefix),
- Qualified Dublin Core (all tags starting with *dcterms:* prefix), and
- Europeana Semantic Elements (tags with *europeana:* prefix).

To make the process of indexing more effective the following set of fields was applied in the descriptions: <dc:contributor>, <dc:creator>, <dc:source>, <dc:description>, <dc:date>, <dc:language>, <dc:subject>, <dc:title>, <dc:type>, <dc:identifier>, <dc:rights>, <dcterms:alternative>, <dcterms:created>, <europeana:country>, <europeana:language>, <europeana:type>, <europeana:uri>, <europeana:year>. An example of an object from the Polish collection is presented in this mixed XML format on the Figure 1.

Fig. 1. An example of a record from the Europeana Polish collection

```
<ims:metadata ims:language="pol"
ims:namespace="http://www.europeana.eu/"
ims:identifier="http://www.europeana.eu/resolve/record/92
033/8C6DB268A10A2A74B31547D5484BD35A03906704">
<ims:fields>
<dc:date>[ca 1910]</dc:date>
<dc:format>text/html</dc:format>
<dc:identifier>http://193.59.172.16/szzz/ShowStart.do?id=
21505</dc:identifier>
<dc:language>pol</dc:language>
<dc:publisher>[S.l.] : nakład M. Ozura "Kurlandski
magazyn"</dc:publisher>
<dc:publisher>Zakład Reprografii i Digitalizacji Zbiorów
Bibliotecznych Biblioteki Narodowej, 2008</dc:publisher>
<dc:rights>Biblioteka Narodowa</dc:rights>
<dc:source>Biblioteka Narodowa, Poczt.13491</dc:source>
<dc:subject>Druskienniki (Litwa) -
ikonografia</dc:subject>
<dc:title>Pozdrowienie z Druskienik : brzegi Niemna przy
fermie = Privĕt" iz" Druskenik" : bereg" Nĕmana pri fermĕ
[Dokument ikonograficzny]</dc:title>
<dc:type>pocztówka</dc:type>
<dcterms:alternative>Privĕt" iz"
Druskenik"</dcterms:alternative>
<europeana:country>poland</europeana:country>
<europeana:dataProvider>The National Library of Poland -
Biblioteka Narodowa</europeana:dataProvider>
<euro-
peana:isShownAt>http://193.59.172.16/szzz/ShowStart.do?id
=21505</europeana:isShownAt>
```

```
<euro-
peana:isShownBy>http://193.59.172.16/szzz/IsShownBy.do?id
=21505</europeana:isShownBy>
<europeana:language>pl</europeana:language>
<euro-
peana:object>http://193.59.172.16/szzz/IsShownBy.do?id=21
505</europeana:object>
<europeana:provider>The European
Library</europeana:provider>
<euro-
peana:rights>http://creativecommons.org/publicdomain/mark
/1.0/</europeana:rights>
<europeana:type>IMAGE</europeana:type>
<europeana:uri>http://www.europeana.eu/resolve/record/920
33/8C6DB268A10A2A74B31547D5484BD35A03906704</europeana:ur
i>
<europeana:year>1910</europeana:year>
</ims:fields>
</ims:metadata>
```

3.2 Topics

Fifty queries were created so as to represent proportional shares of common names, named entities (anthroponymes, toponymes, urbonymes, names of historical events etc.) and their combinations. They included in city names (e.g. in #13 *Toruń*), region names (e.g. in #25 *Lubelszczyzna* – Lublin region), persons' names (e.g. in #11 *Józef Bem*), event's names (e.g. in #8 *powstania polskie* – Polish uprisings), professions' names (e.g. in #5 *inżynierzy* – engineers), state names (e.g. in #5, #9, #19, #30 *Polska*) and artifacts names (e.g. # 5 *mosty* – bridges). The topics were prepared on the basis of Europeana search logs and of deductions about users interests concerning Polish cultural heritage. Simple queries were composed of two to five terms (three terms on the average).

Terms were used in singular or in plural, many of them were inflected by case; they included polysemic or homonymic words. For example in the topic *obraz wsi* the noun *obraz* means *picture*, *painting* or *image* in a metaphoric sense; the genitive *wsi* comes from nom. *wieś* and means *of a village* or *of a country*. However, as an abbreviation of *Wojskowe Służby Informacyjne WSI* may also signify *military information service* and produce hits in the resources of the archives of the anticommunist opposition (such as KARTA). These inconsistencies were introduced deliberately in order to simulate real users behaviour who are likely to make mistakes or act by trial and error method. As it was mentioned earlier, verbs appeared rarely in topics descriptions, mostly as present or past participles.

The content of some document fields was delivered also in English or in German in addition to Polish. For example institutions such as universities appear under Polish and English names and scientific documents are provided with short abstracts in English. German appears often in the documents published before 1945 on today's western territories of Poland (Lower and Upper Silesia, Pomerania). For example the city of *Toruń* has a German name *Thorn*, the city of *Wrocław* is called in German *Breslau*.

Each topic was identified by the tag <identifier>, while the basic query was provided within the tag <title>. For each query an additional field <description> was provided. The aim of this field was to give the relevance assessors an idea about polysemic, too general, or unclear topics. As stated in [6] the field <description> was not used for retrieval purposes.

Fig. 2. Example of a Polish topic coded in XML

```
<?xml version="1.0" encoding="UTF-8"?>
<topics>
<topic lang="pl">
<identifier>CHIC-2013-PL-030</identifier>
<title>Polska i Europa w 18 wieku</title>
<description>prace poświęcone związkom lub kontaktom Pol-
ski z państwami europejskimi w 18 wieku</description>
</topic>
```

Fig. 3. Example of a Polish topic in English translation

```
<?xml version="1.0" encoding="UTF-8"?>
<topics>
<topic lang="en">
<identifier>CHIC-2013-PL-030</identifier>
<title>Poland and Europe in the 18th century</title>
<description>works on contacts and relations with Euro-
pean countries in XVIII century</description>
</topic>
```

3.3 Manually enriched topics

In the Wrocław part of the experiment topics were manually enriched by three categories of users: students of information science, experienced professionals of information retrieval, and humanities-oriented educated persons. In this way three series of system responses could be generated and evaluated, apart from the basic one without enrichment [8].

These groups were defined in a way to simulate preferences of real users. Actually Europeana at this stage of its development is not addressed to unprepared, general public. It was thus assumed that persons interested at the cultural heritage of the continent are likely to have some educational or professional background relevant to the field covered by the search topics they use. As the evaluation process of this kind is carried out for the first time on Polish resources and their potential users, there were no prior expectations concerning information strategies of the selected groups.

It turned out that "humanities-oriented educated" users most often enriched topics with linguistic means such as synonyms of the existing terms. Expert users who have more encyclopedic knowledge preferred narrower terms related to facts described by the existing terms. E.g. the topic *Polish uprisings in Kingdom of Poland* elicited at this group names of commanders or battles rather than synonyms of the noun *uprising* (this was typical for the "educated users"). Finally students who have rather limited encyclopedic knowledge (especially about history) and little interest at cultural heritage issues applied mixed heuristic strategies.

Fig. 4. Example of a topic enriched by an expert user

```
<topic lang="en">
<identifier>CHIC-2013-PL-030</identifier>
<title>Poland and Europe in the 18th century</title>
<enrichment>Northern war, partitions of Poland, Russia,
Austria, Prussia, Sweden, France</enrichment>
<description>works on contacts and relations with Euro-
pean countries in XVIII century</description>
</topic>
```

3.4 Pre-indexing strategies

For each enriched file as well as for the collection documents stop word removing procedure was applied. The stoplist consists of 304 most frequent and semantically insignificant entries. The list includes determinants, prepositions, conjunctions, pronouns (with their inflected forms). The stop words removal procedure was applied for manually enriched topics files, as well as for the collection itself. As weighting scheme for all the runs OKAPI (BM25) probabilistic algorithm was used [4, 9, 10].

3.5 Evaluation

Results have been evaluated according to the following evaluation schemes: MAP, P@5, P@10, p-value, GMAP, MFRS. Finally, as for previous CHiC experiments, Mean Average Precision (MAP) values were applied. For each topic MAP value has been computed for the first 1000 retrieved documents in a ranked list.

4 Results and analyses

The analysis of the results led to a series of conclusions, which can be grouped in four categories:

- quality of descriptions
- inconsistencies due to the variety of items and exhibits types covered by Europeana
- language-specific difficulties related to Polish as an IR working language
- information retrieval algorithm implemented in the Europeana search-engine

As far as the quality of descriptions is concerned Europeana does not bear much responsibility for errors or shortages. Exhibits and documents are indexed independently by their host institutions. Some conclusions or suggestions could, however, help these institutions to modify and unify their description templates. In particular the fields containing the most significant semantic information, such as <dc:subject>, <dc:title> and <dcterms:alternative> should be prepared with care, according to a set of coherent principles. These fields often determine the result of a query and user's satisfaction.

The second category of conclusions concerns difficulties related to processing of various items and exhibits types. The engine retrieves, amongst others, books, paintings, photographs of architectural and/or natural objects, films, patent descriptions, scientific papers, press items (journals, periodicals), archived administrative documents, or documents of everyday life (e.g. posters, tickets, etc.) Every document type is described according to its own, often specific format, and every country has its own tradition of creating metadescriptions. This variety of local "information cultures" constitutes the richness of European heritage and makes the Europeana project so meaningful. But this is precisely the point where the search-engine encounters the most important obstacle to the effective information retrieval. The question remains open, how to define the reasonable frontier between indexation standards which should remain specific to the country or institution and those which should be imposed as global.

The third set of conclusions concerns linguistic difficulties related to formal, morphosyntactical and semantic phenomena typical for Polish (as well as for other Western-Slavic languages). Most of them were described in the Section 2. Formal constraints of Polish require respecting all the characters of the alphabet, even if some of them might seem unknown to the users. Morphosyntactical constraints include Polish flexion which in many cases influences word meaning. Finally semantic information may be distorted by word homonymy and polysemy. Evaluation process and quantitative measures obtained in the experiment do not offer clear suggestions concerning NLP pre-processing of the queries. However, the fact that basic queries turned out to be more effective than those with manual enrichment is significant (cf. Table 2). One should stress that in a flectional language like Polish stemming and stop words removal do not necessarily enhance the effectiveness of retrieval [11].

Finally, the fourth category of conclusions concerns the information retrieval algorithm implemented in the Europeana search-engine. The results for all the evaluation measures are presented in the Table 2.

Table 2. Evaluations of the submitted runs.

Run id	MAP	P@5	P@10	p-value	GMAP

Baseline	0.3140	0.5880	0.5520		0.1515
PLWREdu	0.1529	0.4000	0.3500	0.000080	0.0237
PLWRExp	0.1795	0.4200	0.3780	0.006573	0.0507
PLWRStu	0.1279	0.2880	0.2680	0.000020	0.0376

All the average measures yield the highest values for the basic queries. This means that introducing new terms (enrichment) does not improve the results. Such a conclusion is, at least apparently, counter intuitive. Assuming the correctness of the IR algorithm more information given by a user should result in better scores. Indeed, this would be the case if some terms, such as for instance the first and the last name of a person occurring in one query were connected by conjunction. If all the terms are connected by disjunction, as it was the case, every single term enlarges the number and the scope of documents retrieved, many of them being irrelevant. Reliable comparisons may thus be conducted between queries composed of a similar number of search terms. Actually, when enriched runs are compared alone the results correspond to the expectations, because the highest precision of the retrieved information is attained by expert users and the lowest one by students.

If the MAP measure is taken into consideration the decrease of precision is very high between basic and the other queries, but remains low between the enriched queries PLWREdu, PLWRExp and PLWRStu (Table 3).

Table 3. Decrease of the MAP values

Run id	Parameters	MAP	% of change
Baseline	<title> field, no stop words	0.3140	
PLWREdu	educated users, no stop words	0.1529	-51.3%
PLWRExp	expert user, no stop words	0.1795	-42.8%
PLWRStu	student, no stop words	0.1279	-59.3%

5 Conclusions

The experiment shows that working on Polish topics in Europeana is not sufficiently effective and important improvements are necessary in the search engine. There were only 26.6% of highly relevant responses to the queries and 14.8% of partly relevant ones. If partly relevant and irrelevant responses are considered as one set, as much as much as 73.5% of queries were judged by the users as unsatisfactory. Automatic and manually enriched queries were evaluated by users and the results of these evaluations were compared with respect to the Mean Average Precision. MAP produced the best results for the automatic search (0,314), while enriched files contained less relevant results: for the expert users MAP value was 0,1795, for the educated users it was 0,1529 and for students 0,1279.

The results are due to the incoherence of object descriptions and to some linguistic traps of Polish as a retrieval language. An important factor seems to be the algorithm analyzing the entire descriptions without any specification of relevant or irrelevant fields. The problem of sensitivity of the IR systems to the specific Polish characters should be also treated. Finally uniterm indexing strategy, based on matching documents to a single keyword from topic, is not the best choice.

Should Europeana become a truly European project, specific features of the description languages representing a large number of items in the resources ought to be respected, because an integral information system includes also its users and their habits.

References

1. Bielec D.: Polish: an essential grammar. London; New York : Routledge, 2003.
2. *CHiC: Cultural Heritage in CLEF*, http://www.promise-noe.eu/chic-2013/home
3. *Europeana Europeana: think culture*, http://europeana.eu/
4. Fautsch C., Savoy J.: *Algorithmic Stemmers or Morphological Analysis: An Evaluation*. JASIST. 60, 1616-1624 (2009)
5. Feldstein R. F.: *A Concise Polish Grammar*, SEELRC 2001 http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=4
6. *Guidelines for participation and submission*, on-line: http://www.promise-noe.eu/chic-2013/guidelines-for-participation-and-submission/polish-task
7. Jagodzinski G.: *A Grammar of the Polish Language*, http://grzegorj.w.interia.pl/gram/en/gram00.html.
8. Petras V., et all: *Cultural Heritage in CLEF (CHiC) Overview 2012*, on-line: http://www.clef-initiative.eu/documents/71612/0cadb163-3e32-4f16-a659-b457480c2a29
9. *Polish Track at CLEF 2013*, http://members.unine.ch/jacques.savoy/Polish/
10. Robertson S.: How Okapi Came to TREC, in: Harman D., Voorhees Ellen V.: TREC. Experiment and Evaluation in Information Retrieval, (287-299), The MIT Press (2005)
11. Savoy J.: *Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages*. Proceedings ACM-SAC, 1031-1035. The ACM Press, (2006)
12. Swan Oscar E.: *Polish Grammar in a Nutshell*, http://polish.slavic.pitt.edu/firstyear/nutshell.pdf