Corpus Metadata Harmonization

Ismail El Maarouf, Roser Saurí

Oxford University Press, Great Clarendon Street, OX2 6DP, Oxford, UK

E-mail: {Ismail.ElMaarouf, Roser.Sauri}@oup.com

Abstract

This paper addresses a major need in the development of quality standards for language corpora: Corpus Metadata Harmonization (CMH). The idea behind CMH is to provide a single framework for metadata encoding that would validate any corpus against one single specification. Operating with a single framework improves consistency, corpus interoperability, and reuse, and increases control by Corpus Querying Systems (CQS).

CQS are used by lexicographers and language specialists to find evidence for words and patterns in corpora. CQS help them to manipulate metadata to isolate and compare word uses: typical operations include filtering out medical usage, contrasting e.g. British with American English, or finding new words. None of these tasks is possible without metadata (professional domain, geography, and time metadata, respectively), generally defined at the document (or higher) level. However, corpus metadata come in a variety of formats, and degrees of quality and granularity, and corpora are often created for a context-specific purpose. The challenges of CMH are to single out major metadata categories and unify existing resources under a single model while preserving the original metadata related to the original purpose the corpus was created for.

This paper takes a Semantic Web approach to CMH and proposes a cross-corpus ontology for metadata, built up from user requirements (dictionary editors) and major resources available at Oxford University press (OUP). It reports on experiments mapping single-corpus metadata categories to this cross-corpus metadata ontology, and results obtained using the Sketch Engine CQS.

Keywords: corpus metadata harmonization; corpus query system; sketch engine; corpus lexicography; semantic web; ontologies

1. Context

The last few decades have seen a significant surge in the amount of data available to linguists and lexicographers (Ferraresi et al. 2008; Pomikálek et al. 2009), with text corpora reaching tens of billions of words. The quality of such (web) giga-corpora is not satisfactory, but specialized users still find them useful and gradually adjust to noisy data. At OUP, the Dictionary Division has long built large corpora for lexicography purposes (revising entries, collecting evidence, harvesting new words) while putting an emphasis on quality. The Oxford English Corpus (OEC), for instance, was created by manually defining metadata from each source website.

Despite the fact that OUP corpora (currently mainly for English; see table 1) have been produced within different contexts using diverse sources, substantial efforts have been put into processing them using a single framework. For example, they uniformly use the same part of speech tagset and are all processed with the same NLP (Natural Language Processing) pipeline. This ensures that the data model is consistent across corpora, and allows corpus users to make accurate comparisons. Efforts are now directed towards Harmonizing Corpus Metadata (CMH) information, for example information on the domain of a text or on its date of publication. This is crucial for the following reasons:

* *Enable cross-corpus search:* CMH enables tasks, like filtering out medical usage, contrasting British and American English, and finding new words, which rely on metadata (professional domain, geography, and time metadata, respectively), to be performed across corpora.
* *Facilitate resource interoperability:* CMH opens up a new range of applications, revolving around on-demand corpus services; e.g. building dedicated corpora from a variety of interoperable sources, or calibrating corpus input for the production of corpus-derived datasets (e.g. wordlists).
* *Increase robustness in corpus creation:* CMH provides a single framework to validate any corpus resource against one single metadata specification and thus improve quality in corpus delivery.

Metadata harmonization is a difficult task as each corpus comes with its own classification, so two categories with the same name do not necessarily share the same definition. Moreover, classifications differ in terms of granularity, where for example, a particular category subsumes a rich hierarchy because the corpus contains a large number of corresponding documents. It is often necessary to review the texts of a particular category in order to understand its use. This method has its limits as (1) clear-cut boundaries can be hard to set, and (2) it is impossible to review metadata categories for giga-corpora (sampling or clustering algorithms can be used).

This paper reports on current efforts at OUP in CMH, and is organized as follows: section 2 provides a description of the model and methodology, section 3 describes experiments on the benefits of harmonized corpora for New Word Detection and section 4 concludes the paper.

1. Model and methodology

The metadata model

CMH involves defining ontologies for existing metadata, and mapping them to a cross-corpus ontology. An analysis of the largest corpora available at OUP Dictionaries produced a list of top-level metadata categories or main types of information needed to be encoded with a corpus. These are: **Time, Location, Source, Domain, Genre,** and **Language**. Each of these top-level categories subsumes several categories. *Location* and *Time* are relatively straightforward to define and to populate and they can easily be structured into hierarchies. *Source* information relates both to publisher and to author information, and *Language* should convey information about language (English, French), the mode (written, oral, etc.) and register (technical, formal, slang, etc.). *Domain* generally specifies the topic of a text and *Genre* relates to the text type or discourse mode. Each top-level category was defined in a cross-corpus metadata ontology that was linked to single corpus metadata ontologies via mappings. Once these models are in place, it is possible to make corpora interoperable, so as to implement cross-corpus search and corpus validation solutions. Metadata mappings have so far only been completed for *Time* and *Location*. *Time* and *Location* are both structured into hierarchies: *Time* is structured into *Century*, *Decade*, *Year*, and *Month*; *Location*, into *Region (area larger than a country; e.g. North America)*, *Country*, and *Division (area smaller than a country; e.g. New England)*.

1. The Oxford Corpus Portfolio

We used the virtual corpus feature of the Sketch Engine (Kilgarriff et al., 2004), which enables collection of part or all of several corpora into one “metacorpus”, in order to provide access to the complete set of harmonized corpora, the Oxford Corpus Portfolio (OCP), worth 13 billion tokens, and combining diverse sources (see Table 1).

|  |  |  |
| --- | --- | --- |
| **Corpus** | **Description** | **Size** |
| Oxford Corpus of Academic English - Journals (OCAEJ) | *OUP Journals indexed with the Oxford Taxonomy.* | 1.6 |
| Oxford English Corpus (OEC) | *Web content from quality web sites as well as print content such as books and journals.* | 2.5 |
| New Monitor Corpus(NMC) | *Web content from twice-daily trawls of RSS feeds.* | 8 |
| Oxford Twitter Corpus(OTC) | *Social media content from Twitter API (fixed, dated 2012).* | 0.1 |
| Corpus of Historical American English (COHA) | *Balanced historical corpus of American English[[1]](#footnote-1)* | 0.4 |

Table 1 – Corpora federated in the Oxford Corpus Portfolio (size in billions of words)

Figure 1 – Time span coverage of OCP corpora

Within the Sketch Engine, the interface of the OCP works like any other corpus. Compiling a virtual corpus saves time compared to compiling a corpus from scratch, even when this requires, as it was the case, adding new cross-corpus attributes to each corpus. Figure 1 illustrates the time span covered by each corpus. The red blocks indicate spans where more than one corpus overlaps.

1. Evaluation
2. New Word Detection

The aim of the experiment was to assess whether harmonized corpora improve the quality of New Word candidate lists automatically extracted from metadata-controlled subcorpora from OCP and NMC (New Monitor Corpus). The OCP combines NMC with a large balanced corpus (OEC), an academic corpus (OCAEJ), and a historical corpus (COHA), so provides richer contexts for words.

To generate new word candidates, we applied the keyword extraction feature of the Sketch Engine (see Kilgarriff et al 2009), which ranks words according to their frequency in a *focus* corpus and in a *reference* corpus (which is typically bigger). Specifically, the algorithm computes a ratio of the relative frequency of a word in each corpus as expressed in the following equation:

$$Ratio\_{i}= \frac{F\_{i,fc }.\frac{1}{Size\_{fc}}+∝}{F\_{i,rc }.\frac{1}{Size\_{rc}}+∝}$$

*Where fc is focus corpus, rc is reference corpus and Fi, rc is the frequency of wordi in rc. The ∝ parameter was set to 1 to handle out-of-vocabulary words (words unseen in the reference corpus). Values are positive, there is no maximum, and low values indicate that the frequency of words is similar in the fc and in the rc.*

We hypothesize that in the case where we have more evidence for a word, we should expect its *ratio* to decrease, making it less likely to be a new word. We applied the algorithm to extract American new words and British new words from 2016 in NMC and in OCP. The experiment amounts to keeping the same focus corpus (US-2016 or UK-2016) and only changing the reference corpus (respectively, OCP-US and NMC-US for US-2016, or OCP-UK and NMC-UK for UK-2016). Parameters used were:

* The size of the list is 1000 words;
* The minimum frequency for a word in the focus corpus is 5;

The word forms are filtered according to a regular expression, to exclude proper nouns, badly tokenised strings, and plurals: ^[a-z][a-zA-Z\-]\*[^s]$

1. Automatic comparison against dictionary sources

The resulting ranked word lists were combined to identify duplicate words and matched against the lemmas and inflections of the following OUP dictionaries:

* ODE (Oxford Dictionary of English, 02/2016): 359,860 entries
* NOAD (New Oxford American Dictionary, 02/2016): 336,199 entries
* OED (Oxford English Dictionary, 04/2016): 931,881 entries

The words were also checked for their presence in Lemur, the OUP database which contains suggestions for new words to be added to dictionaries – words spotted by editors. Each word was therefore complemented with variables storing (boolean) values indicating their presence in a given source.

Pairwise correlations were computed to analyse whether these variables tended to follow a common pattern, and significance testing (T-test) was also computed to see if statistics were reliable. In Table 2, a large portion of correlation values are very significant (with two stars; 0.01>p> 0.001); weak (absolute values between 0.2 and 0.39; in orange), moderate (absolute values between 0.4 and 0.59; in blue), and strong (absolute values over 0.6 in green) correlation values have been observed.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **NMC-UK** | **NMC-US** | **OCP-UK** | **OCP-US** | **ODE** | **NOAD** | **OED** | **LEMUR** |
| **NMC-UK** | 1 | -0.31\*\* | 0.42 | -0.45 | -0.02 | -0.09 | -0.01\*\* | -0.06\*\* |
| **NMC-US** |  | 1 | -0.26\*\* | 0.40\*\* | -0.05\*\* | 0.04 | -0.08\*\* | 0.09\*\* |
| **OCP-UK** |  |  | 1 | -0.29 | 0.04 | -0.05 | 0.07\*\* | -0.02\*\* |
| **OCP-US** |  |  |  | 1 | -0.06 | 0.04 | -0.01\*\* | 0.14\*\* |
| **ODE** |  |  |  |  | 1 | 0.80\*\* | -0.08\*\* | 0.11\*\* |
| **NOAD** |  |  |  |  |  | 1 | -0.33\*\* | 0.20\*\* |
| **OED** |  |  |  |  |  |  | 1 | -0.03\*\* |
| **LEMUR** |  |  |  |  |  |  |  | 1 |

Table 2- Correlation and significance testing

It is interesting to note that correlations between US and UK dialects are all negative, which reflects the fact that they tend to produce different word lists (as expected when controlling location metadata), although correlations are weak to moderate (-0.31\*\*, -0.45, -0.26\*\*, -0.29), and only NMC-UK/NMC-US, OCP-UK/NMC-US are significant. ODE is very strongly (positively) correlated with NOAD (significant), which means that if a word is in ODE, it is highly likely to be in NOAD and vice versa. The main reason they are correlated is that NOAD is based on ODE, and most of what is added to one is added to the other. This is not the case for OED, which is negatively correlated with NOAD (-0.33\*\*), and not correlated to ODE (-0.08\*\*): one reason is the difference in editorial policies. There were no correlations observed between corpus lists and dictionary sources, because a large portion of new word candidates are already in dictionary sources.

1. Manual evaluation

We combined candidates from each wordlist and automatically removed those present in both ODE and NOAD at the same time (equivalent to a blacklist filter), reducing the candidate list to a pool of 258 unique candidates. This was manually analyzed by one editor. 6% of words consisted of errors (e.g. due to bad tokenization, or HTML tags), the rest, of new material: *new variants* (valid variant word form of an existing headword), *new words* (headwords absent from a dictionary source –not necessarily “new” in the sense of recently created) or *new senses* (senses absent from an existing word entry), depending on the source dictionary. New words consisted of 75% or more of the filtered lists, as seen in Table 3.

|  |  |  |  |
| --- | --- | --- | --- |
| **corpus** | **correct** | **total** | **accuracy** |
| **OCP-US** | 91 | 106 | 0.86 |
| **NMC-UK** | 92 | 112 | 0.82 |
| **NMC-US** | 60 | 78 | 0.77 |
| **OCP-UK** | 77 | 102 | 0.75 |
| **NMC** | 128 | 160 | 0.80 |
| **OCP** | 138 | 172 | 0.80 |

Table 3 – New word accuracy

Table 3 indicates that scores are equivalent overall, but that NMC is more suited for UK new words, while OCP is better for US new words. From a lexicographer point of view, it is also interesting that OCP identified c.8% more words than NMC alone without any loss of accuracy.

1. Example analysis

The new word candidates identified in the previous subsection were further classified according to whether they were *low* or *high* priority new words. This decision is based on various criteria, such as morphology, semantic compositionality, productivity of a compound pattern, frequency, terminological domain, or the identification of a gap in one dictionary source. 88 words were classed as high priority and 117 words were classed as low priority. Examples of high-priority words include *academisation*, *ride-hailing*, or *undrafted*, and examples of low priority words included transparent compound adjectives such as *anti-EU*, *plant-based*, *or anti-Trump.*

Figure 2 – Breakdown of manual evaluation by corpus and category

We computed the proportion of high and low priority words as well as spelling variants and errors for each subcorpus and each corpus. Figure 2 reveals that OCP-UK proportionally helps to extract fewer low priority new words for a similar amount of new words, compared to NMC-UK. Similarly NMC-US collects proportionally fewer new words than its corresponding OCP subcorpus while extracting more errors and variants. The results seem to suggest that CMH could help in better prioritizing/ranking new words. This is confirmed when comparing NMC and OCP regardless of location: OCP collects proportionally more new words and fewer low priority words for an equal number of errors and slightly higher number of variants.

1. Conclusions

Corpus Metadata Harmonization is an ongoing project at OUP which provides a unified framework for improved consistency, validation, and interoperability of corpus resources. This paper provided the description of a model and methodology to implement a CMH solution by mapping existing corpus metadata to a cross-corpus ontology. It also reported on results obtained with the implementation of a cross-corpus search service on the Sketch Engine. Finally it described an experiment assessing how harmonized metadata corpora benefits new word detection. On this last matter, it is probably too early to draw conclusions, as more experiments need to be performed to get a better insight of the algorithm used, and understand how CMH contributes to New Words Detection. However, this paper has shown that new words can be extracted accurately with relatively simple methods. Future work includes

* Using rank scores from the algorithm to prioritize new words; this implies defining priority categories as suggested in section 3.4.
* Comparing different ranking algorithms for new words detection.
* Producing harmonized metadata for categories other than Time and Location.
* Applying the metadata model to corpora in languages other than English
1. Acknowledgements

We would like to thank the Editorial Content Group at OUP for their input and feedback, particularly Judy Pearsall, Angus Stevenson, Katherine Martin, Jeffrey Sherwood and Rebecca Hotchen.

1. References

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating "ukwac", a very large web-derived corpus of English. In Proc. *WAC4 Workshop at LREC*, Marrakech, Morocco.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proc. Corpus Linguistics*, Liverpool.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud*, pp. 105–116.

Pomikálek, J., Rychlý, P., and Kilgarriff, A. (2009). Scaling to billion-plus word corpora. In *Advances in Computational Linguistics: Special Issue of Research in Computing Science*, volume 41, Mexico City.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus*. Gedit.

Sinclair, J. (1991). *Corpus, concordance, collocation: Describing English language*. Oxford: Oxford University Press.

**Dictionaries:**

ODE: Oxford Dictionary of English. Oxford: Oxford University Press.

NOAD: New Oxford American Dictionary. New York: Oxford University Press.

OED: Oxford English Dictionary. Oxford: Oxford University Press.

**Corpus:**

Davies, Mark. (2010-) The Corpus of Historical American English: 400 million words, 1810-2009. Available online at http://corpus.byu.edu/coha/.

This paper is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



1. This corpus is licensed from BYU (Davies, 2010) [↑](#footnote-ref-1)