Visual Comparison of Multilingual Documents and Lexical Matching

Minoru Nakayama [(1)], Mountaz Hascoët [(2)]
(1) Human System Science / CRADLE, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo, Japan
(2) LIRMM, UMR 5506 du CNRS, Univ. Montpellier II, 161, rue Ada 34392 Montpellier Cedex, France
nakayama@cradle.titech.ac.jp, mountaz@lirmm.fr

Abstract

In this paper, we propose a new approach for the comparison and matching of documents written in different languages. Our approach is devoted to situations where documents can be conceptually represented by abstract graphs. Our approach builds on previous work yet it addresses some limitations left open by previous work. The benefits of our approach is twofold. First, our approach makes it possible to compare original multi-lingual documents without requiring a translation in a pivot language contrary to other approaches. Second, we propose a highly visual and interactive environment so that human experts can perform both comparison of documents and lexical matching in a seamless way contrary to other approaches where matching and comparison are often considered separately.

Keywords- Multi-Layer Graph Matching and Comparison, Heat Map based Visual Comparison, Multi-lingual documents.

1. Introduction

There are many requests to compare the documents across multiple languages. A typical approach starts with an automatic translation via English, as for example for patent documents. Most automatic translation researches are based on the translation corpora between two languages which are called as parallel corpora, some procedures of extracting translation pairs of words have been studies [2]. These techniques provide the benefits which need not specialized dictionary, word pairs for any language combinations can be easily created [3]. Though many methodologies of creating pairs of word translations have proposed since the mid of 1990s, the performances are always limited [4], [5].

Recently, some development procedures for specific dictionaries and new approaches such as statistical machine translation have been studied [6].

Another approach is to compare the word relational networks based on word co-occurrence across two language sets. The co-occurrence relationship can be illustrated as graphs, the translated relationship can be considered [7]. When the structure of graphs are complicated, the graphs should be clustered and set levels of hierarchies.

However, all these automatic analyses cannot provide appropriate results of translation without some interactive sessions with a human needed to adjust otherwise error prone results.

In this paper, we make the assumption that (1) a graph can be computed to conceptually represent documents in any language and (2) a similarity matrix can be provided as the outcome of automatic multi-lingual lexical analysis to represent computed similarities between pairs of lexical expression in two different languages.

Based on these hypotheses, we build on previous work and in particular on a visual approach proposed for general graph matching and comparison [9], to propose a new approach to the analysis of multi-lingual document collections. However, work in [9], based on invariant layout are limited to cases where graph matching can be expressed as a set of explicit criteria that can be used to produce invariant layouts.

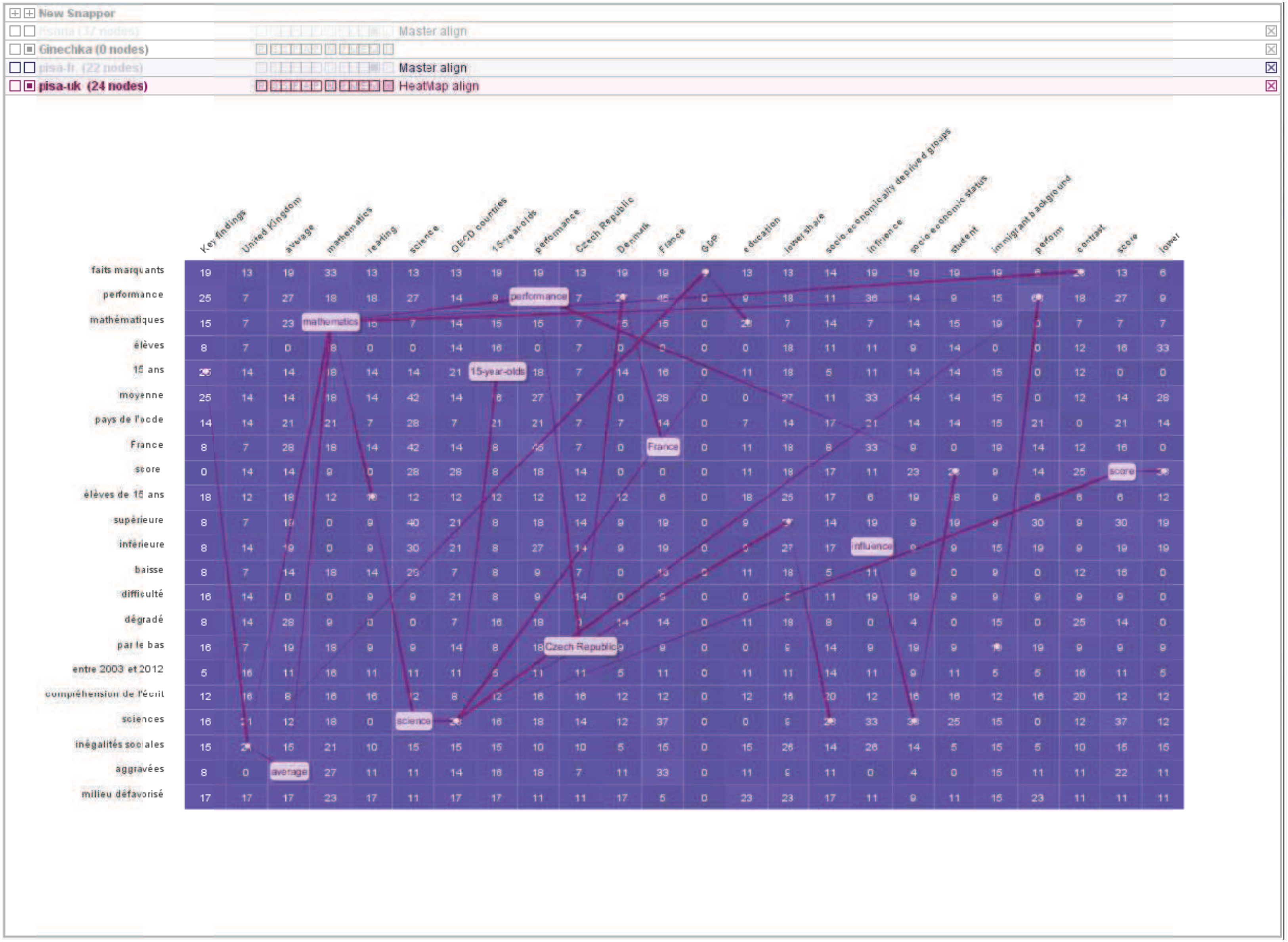
Cases where no explicit criteria are provided with the data to produce invariant layout are relatively frequent. However in these cases, a similarity matrix is often given. Accounting for a similarity matrix was beyond the scope of work in [9]. Our proposal in this paper is to extend previous work to handle these cases by introducing a heat map layer to account for similarity matrices.

We believe that the benefits of our approach is twofold. First, our approach makes it possible to compare original multi-lingual documents directly, without requiring a translation in a pivot language contrary to other approaches nor requiring any alignment of texts. Second, we propose a highly visual and interactive environment so that human experts can perform both comparison of documents and lexical matching in a seamless way contrary to other approaches where matching and comparison are often considered separately.

In this paper, we start by an overview of the proposed approach and further illustrate it through a case study involving English and French documents. We finally conclude with perspectives and future work.

2. Methodology

The methodology proposed in this paper makes two strong assumptions: (1) graphs can be used as skeletal representations of document contents, and (2) relationships between lexical expressions of different languages can be expressed by a similarity matrix, a contingency table or a bipartite graph.



⊞ ⊞ New Snapper

□ □ Psoria (37 nodes) Master align ☒

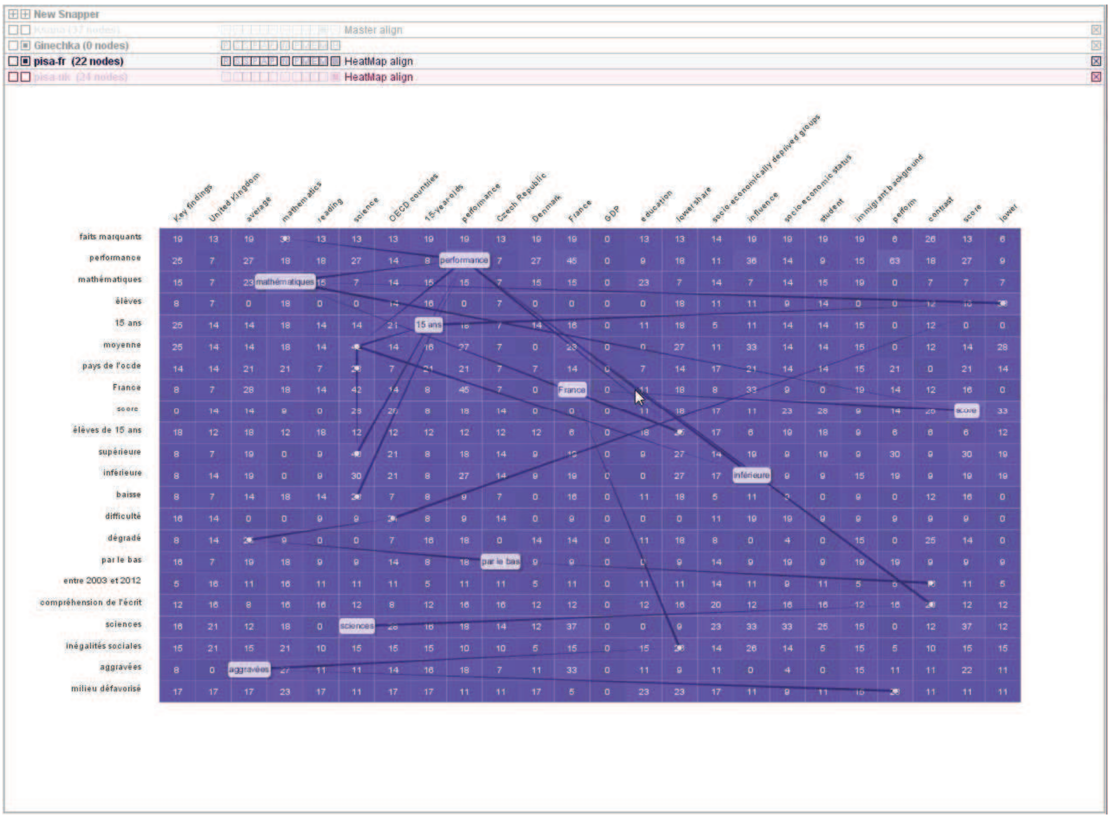
□ ▣ Ginechka (0 nodes) ☒

□ □ pisa-fr (22 nodes) Master align ☒

□ ▣ pisa-uk (24 nodes) HeatMap align ☒

	Key findings	United Kingdom	average	mathematics	reading	science	OECD countries	15-year-olds	performance	Czech Republic	Denmark	France	GDP	education	lower share	socio-economically deprived groups	influence	socio-economic status	student	immigrant background	perform	contrast	score	lower
faits marquants	19	13	19	33	13	13	13	19	19	13	19	19	9	13	13	14	19	19	19	19	6	25	13	6
performance	25	7	27	18	18	27	14	8	performance	7	27	45	0	9	18	11	36	14	9	15	68	18	27	9
mathématiques	15	7	23	mathematics	15	7	14	15	15	7	15	15	0	20	7	14	7	14	15	19	0	7	7	7
élèves	8	7	0	8	0	0	14	16	0	7	0	0	0	0	18	11	11	9	14	0	0	12	16	33
15 ans	25	14	14	18	14	14	21	15-year-olds	18	7	14	16	0	11	18	5	11	14	14	15	0	12	0	0
moyenne	25	14	14	18	14	42	14	16	27	7	0	28	0	0	27	11	33	14	14	15	0	12	14	28
pays de l'ocde	14	14	21	21	7	28	7	21	21	7	7	14	0	7	14	17	21	14	14	15	21	0	21	14
France	8	7	28	18	14	42	14	8	45	7	0	France	0	11	18	8	33	9	0	19	14	12	16	0
score	0	14	14	9	0	28	28	8	18	14	0	0	0	11	18	17	11	23	28	9	14	25	score	38
élèves de 15 ans	18	12	18	12	18	12	12	12	12	12	12	6	0	18	25	17	6	19	18	9	6	6	6	12
supérieure	8	7	18	0	9	40	21	8	18	14	9	19	0	9	27	14	19	9	19	9	30	9	30	19
inférieure	8	14	19	0	9	30	21	8	27	14	9	19	0	9	27	17	influence	9	9	15	19	9	19	19
baisse	8	7	14	18	14	25	7	8	9	7	0	13	0	11	18	5	11	9	0	9	0	12	16	0
difficulté	16	14	0	0	9	9	21	8	9	14	0	9	0	0	0	11	19	19	9	9	9	9	9	0
dégradé	8	14	28	9	0	0	7	16	18	0	14	14	0	11	18	8	0	4	0	15	0	25	14	0
par le bas	16	7	19	18	9	9	14	8	18	Czech Republic	9	9	0	0	9	14	9	19	9	19	19	9	9	9
entre 2003 et 2012	5	16	11	16	11	11	11	5	11	11	5	11	0	11	11	14	11	9	11	5	5	16	11	5
compréhension de l'écrit	12	16	8	16	16	12	8	12	16	16	12	12	0	12	16	20	12	16	16	12	16	20	12	12
sciences	16	21	12	18	0	science	28	16	18	14	12	37	0	0	9	28	33	38	25	15	0	12	37	12
inégalités sociales	15	24	15	21	10	15	15	15	10	10	5	15	0	15	26	14	26	14	5	15	5	10	15	15
aggravées	8	0	average	27	11	11	14	16	18	7	11	33	0	11	9	11	0	4	0	15	11	11	22	11
milieu défavorisé	17	17	17	23	17	11	17	17	11	11	17	5	0	23	23	17	11	9	11	15	23	11	11	11

Figure 1-a Heat Map based layout of English version of the compared documents.



⊞⊞ New Snapper
□□ Ksana (37 nodes) Master align ☒
□▣ Ginechka (0 nodes) ☒
□▣ pisa-fr (22 nodes) HeatMap align ☒
□□ pisa-uk (24 nodes) HeatMap align ☒

	Key findings	United Kingdom	average	mathematics	reading	science	OECD countries	15-year-olds	performance	Czech Republic	Denmark	France	GDP	education	lower share	socio-economically deprived groups	influence	socio-economic status	student	immigrant background	perform	contrast	score	lower
faits marquants	19	13	19	38	13	13	13	19	19	13	19	19	0	13	13	14	19	19	19	19	6	26	13	6
performance	25	7	27	18	18	27	14	8	performance	7	27	45	0	9	18	11	36	14	9	15	63	18	27	9
mathématiques	15	7	23	mathématiques	15	7	14	15	15	7	15	15	0	23	7	14	7	14	15	19	0	7	7	7
élèves	8	7	0	18	0	0	14	16	0	7	0	0	0	0	18	11	11	9	14	0	0	12	16	38
15 ans	25	14	14	18	14	14	21	15 ans	16	7	14	16	0	11	18	5	11	14	14	15	0	12	0	0
moyenne	25	14	14	18	14	42	14	16	27	7	0	23	0	0	27	11	33	14	14	15	0	12	14	28
pays de l'ocde	14	14	21	21	7	28	7	21	21	7	7	14	0	7	14	17	21	14	14	15	21	0	21	14
France	8	7	28	18	14	42	14	8	45	7	0	France	0	11	18	8	33	9	0	19	14	12	16	0
score	0	14	14	9	0	23	26	8	18	14	0	0	0	11	18	17	11	23	28	9	14	26	score	33
élèves de 15 ans	18	12	18	12	18	12	12	12	12	12	12	6	0	18	26	17	6	19	18	9	6	6	6	12
supérieure	8	7	19	0	9	40	21	8	18	14	9	19	0	9	27	14	19	9	19	9	30	9	30	19
inférieure	8	14	19	0	9	30	21	8	27	14	9	19	0	0	27	17	inférieure	9	9	15	19	9	19	19
baisse	8	7	14	18	14	28	7	8	9	7	0	16	0	11	18	5	11	0	0	9	0	12	16	0
difficulté	16	14	0	0	9	9	21	8	9	14	0	9	0	0	0	11	19	19	9	9	9	9	9	0
dégradé	8	14	28	9	0	0	7	16	18	0	14	14	0	11	18	8	0	4	0	15	0	25	14	0
par le bas	16	7	19	18	9	9	14	8	18	par le bas	9	9	0	0	9	14	9	19	9	19	19	9	9	9
entre 2003 et 2012	5	16	11	16	11	11	11	5	11	11	5	11	0	11	11	14	11	9	11	5	5	16	11	5
compréhension de l'écrit	12	16	8	16	16	12	8	12	16	16	12	12	0	12	16	20	12	16	16	12	16	20	12	12
sciences	16	21	12	18	0	sciences	26	16	18	14	12	37	0	0	9	23	33	33	25	15	0	12	37	12
inégalités sociales	15	21	15	21	10	15	15	15	10	10	5	15	0	15	26	14	26	14	5	15	5	10	15	15
aggravées	8	0	aggravées	27	11	11	14	16	18	7	11	33	0	11	9	11	0	4	0	15	11	11	22	11
milieu défavorisé	17	17	17	23	17	11	17	17	11	11	17	5	0	23	23	17	11	9	11	15	28	11	11	11

Figure 1-b Heat Map based layout of French version of the compared documents.

The main important asset of these three analogous representations is that for any expression α of language A and γ of language G, a similarity between α and γ can be expressed and valued, usually with a value comprised between O and 1 (or similarly from 1 to 100). Even though differences exist between similarity measures, bipartite graphs and contingency tables, in the rest of this paper, we will make no distinction between these representation mainly because they serve the same purpose in our methodology

3. Visual Matching and Comparison

Based on the two main assumptions mentioned above, the problem of the analysis of documents in different languages can be seen as a particular case of graph matching and comparison.

A visual approach to graph matching and comparison has been proposed in [9]. Compared to previous work, one of the main assets of the approach of [9], is to integrate the matching and comparison operations in the same interaction model.

Integrating the matching and comparison operations in the same interaction model is achieved thanks to invariant graph layout and multi-layer visualization. Our present work is based on [9] and extend it as will be shown in the next section.

In [9], graphs to be compared are displayed on translucent layers. Displaying graphs on top of each other further facilitate the matching and the comparison provided that relevant layouts automatically produces visual matching of nodes such as in Figure 3. In [9], invariant layouts of graph were proposed for this purpose. However, invariant layout are limited to cases where matching can be expressed as a set of explicit criteria that can be used to produce invariant layouts.

In other cases, similarities can be computed. Our proposal in this paper is to use a heat map to handle these cases.

4. Heat map: Visualizing Similarities

A heat map is a graphical representation of data where colors are used to display values. When a similarity matrix between expressions in one language and expressions in another language can be computed, a heat map can be used not only to display these values but also serve as a layout strategy for the graphs to be compared. In our approach, the heat map is displayed in the bottom layer making it possible for several graphs to be position on top of it.

A heat map can be considered also as the graphical representation of a valued bipartite graph. Such a bipartite graph can be computed for any pair of graphs provided that nodes can be compared using a similarity measure. In our case, as mentioned earlier we used a Levenshtein based similarity measure but other similarity measure could be favored. In any cases, once the bipartite graph is generated, its heat map representation is straightforward a set of node is displayed vertically and the other set of nodes is displayed horizontally. Similarity values are displayed at the intersection of the y coordinate of horizontally displayed nodes and the x coordinate of vertically displayed nodes.

Layout of nodes is further simple. For a node n in the graph, x and y are computed from the heat map. If n correspond to a vertically (resp. horizontally) displayed node, its x-coordinate is the same as its corresponding node and its y-coordinate (resp. x-coordinate) is the coordinate of the hotspot of the corresponding column (resp. line).

5. Methodology summary

Figure 4 summarizes the most important steps of the methodology introduced in this paper. The multi-layer graph representation makes possible the rapid visualization of documents content or structure, the heat map based layout makes possible the semi-automatic matching of lexical expressions of different languages and finally master graph representation enable a user to save a given matching. Master graph can further be used to layout graphs and maintain matching nodes at the same positions in all subsequent layout

6. Case study

The Pisa reports [13] served as the documents used for the case study. These reports give the results of the Pisa analysis of educations systems in OECD countries. The section key findings of these reports are structured around similar questions, however results language differ since country key results are described in the country language.

The case study consisted in (1) displaying each graph representing each report according to a heat map computed from the expression extracted from the documents and a similarity measure between French and English expression, (2) adjusting the matching resulting from error-prone results leaded by the similarity measure, (3) creating a master graph from the correct match and (4) analyzing resulting matched graphs.

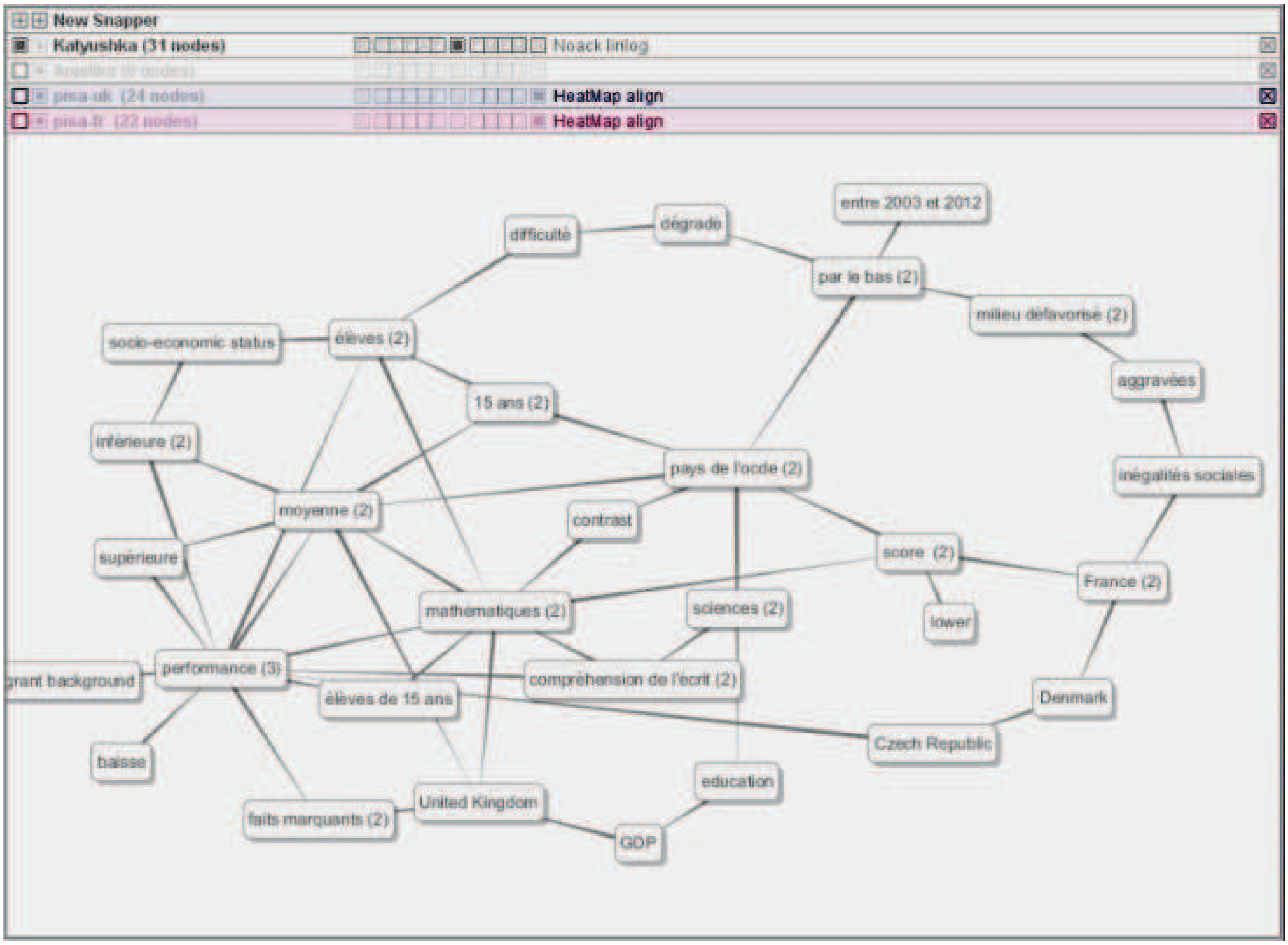


Figure 2 Master graph representing matched English and French documents

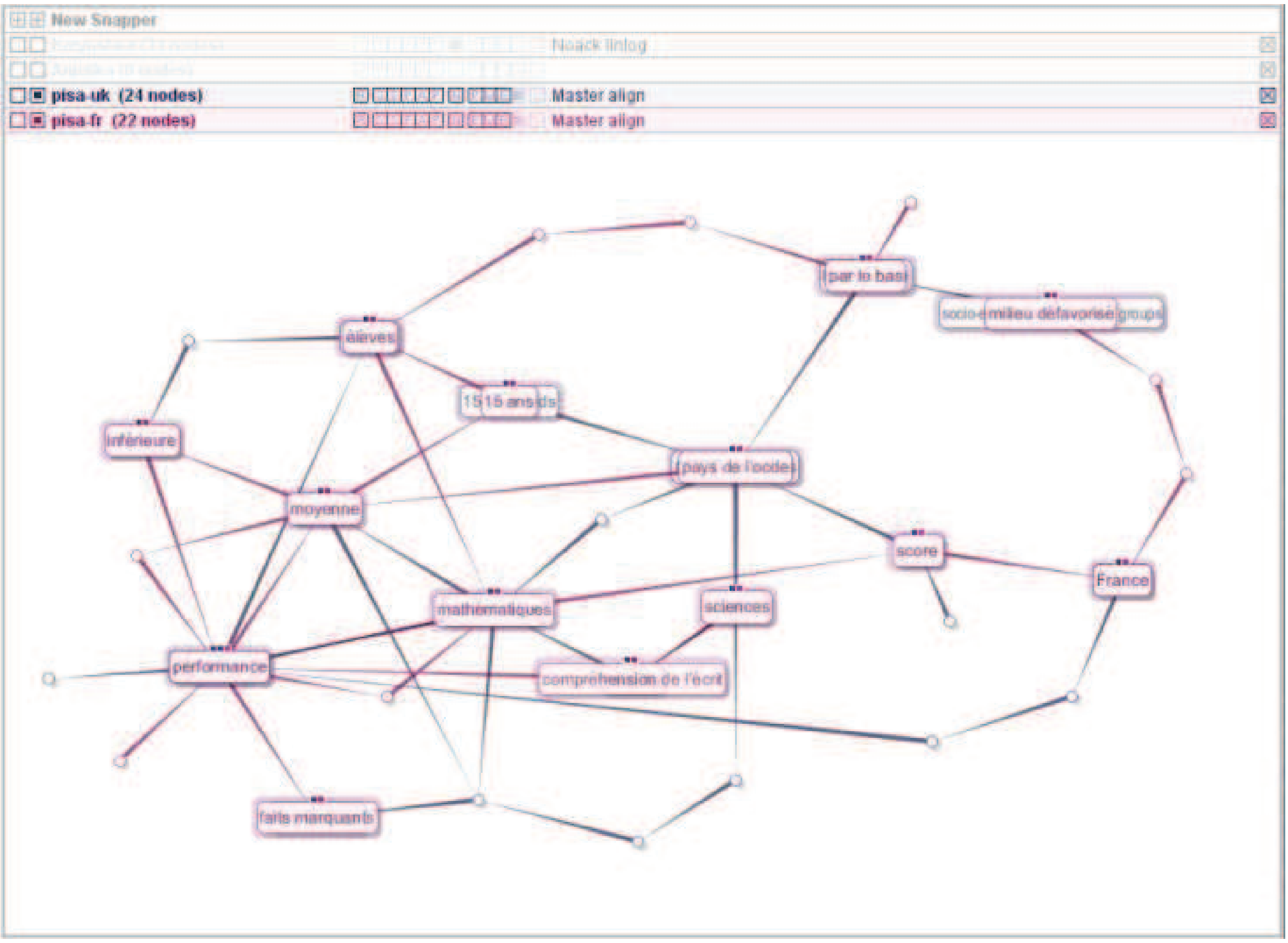


Figure 3 Superposed graphs representing matched English (bottom layer) and French (top layer) documents laid out thanks to the master graph of Figure 2.

Lexical expression extraction can be performed automatically. However, the effort needed in order to extract compound expressions in any language is still costly and since it was not the first purpose of this paper, expressions were extracted by human for each document. From the extracted expression, simple graphs were computed according to simple heuristic: nodes are created for each distinct expressions and an edge is automatically created between two nodes if their corresponding expression appear in the same paragraph.

There are many ways to compute a similarity between lexical expressions in different languages. For the purpose of the case study we chose a straight forward measure based on an edit distance. Since English and French languages belong to the same family of languages, an edit distance has the advantage of performing not too bad, at very low cost. Levenshtein or edit distance [16] consists in computing "the minimal number of insertions, deletions and substitutions to make two strings equal." The distance is symmetric, and it holds $0 \leq d(\alpha, \gamma) \leq \max(|x|,|y|)$. Therefore the similarity between expression α of language A and expression γ of language G, is computed as follows:

$$s(\alpha,\gamma) = 100 \times \frac{d(\alpha,\gamma)}{\max(|\alpha|,|\gamma|)}$$

where $d(\alpha, \gamma)$ represents the Levenshtein edit distance between α and γ. Even though such a similarity based on lexical morphology cannot capture anything else than lexical morphology similarities. In order to measure the relationships between two language an edit distance performance are minimal and other approach should be favored. However, automatically determining the relationships between any expression in any language is still error prone even with tremendous efforts. In the case of this paper as far as two close languages are concerned, edit distance performances were sufficient to illustrate the rest of our methodology and the purpose of our methodology is to make it possible to rapidly correct errors.

Figure 1 displays the results for French (Figure 1-a) and English (Figure 1-b) graph representations of the first section of the reports analyzed. In these screen captures, only the matching terms are displayed with a label corresponding word or expression. Other nodes are displayed as small circles. These views make it easy to visually detect matching errors and adjust them rapidly. Two different types of errors are frequent: wrong match like for example, the English term *average* matched to the French term *aggravées* instead of being matched to term *moyenne*; and mismatch like French expression, *faits marquants*, not being matched to the corresponding expression, *key findings*. All these errors can be corrected simply by (1) superposing the layer displaying each graph and (2) moving the corresponding nodes to either superpose them on top of each other to indicate a match or to move them apart from each other to indicate no match and undo a mismatch.

Once the matching is considered correct is can be saved in a master graph. This master graph is displayed in its own layer and can applied a force layout [17].

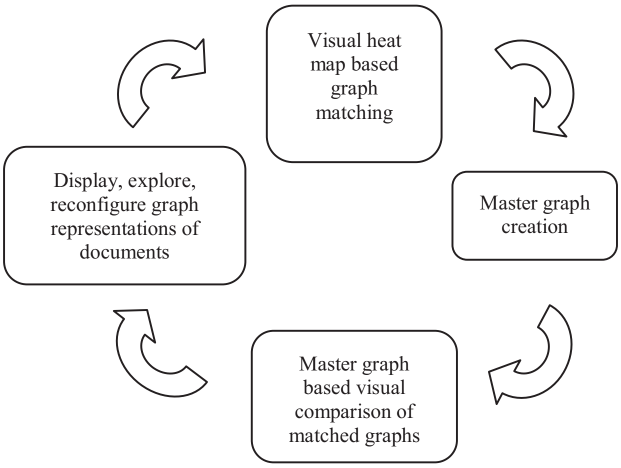

Visual heat map based graph matching
Master graph creation
Master graph based visual comparison of matched graphs
Display, explore, reconfigure graph representations of documents

Figure 4 summary of visual lexical matching and document comparison most important steps

7. Conclusion

The hypotheses taken by our approach covers specific contexts of document comparison and can be considered as a useful complement to other approaches limited to other contexts. Future work may explore the integration in seamless way of visualization of documents coupled with their graph representations.

Another potential perspective for this methodology is to apply it to other application domains, since it is general enough to encompass any other situation where graph representation of data needs to be matched and compared on the basis of some similarity measure such as in biology or chemistry.

8. References

[1] EPO: European patent office. [Online]. Available: http://www.epo.org/searching/free/patent-translation.html

[2] H. Somers, "Review article: Example-based machine translation," Machine *Translation*, vol. 14, no. 2, pp. 113–157, 1999.

[3] F. Smadja, K. R. McKeown, and V. Hatzivassiloglou, "Translating collocations for billingual lexicons: A statistical approach," *Computational Linguistics*, vol. 22, no. 1, March 1996.

[4] P. Fung, "A pattern matching method for finding noun and proper noun translations form noisy parallel corpora," in *ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. ACL, 1995, pp. 236–243.

[5] ——, "Compiling bilingual lexicon entries from a non-parallel english chinese corpus," in *Proceedings of the Third Workshop on Very Large Corpora*, 1995.

[6] A. Lopez, "Statistical machine translation," *ACM Computing Surveys*, vol. 40, no. 3, p. Article 8, August 2008.

[7] K. Tanaka and H. Iwasaki, "Extraction of lexical translations from non-aligned corpora," in *COLING: 16th International Conference on Computational Linguistics*, 1996, pp. 580–585.

[8] G. Artignan, M. Hascoët, and M. Lafourcade, "Multiscale visual analysis of lexical networks," in *Proceedings of IV2009*, 2009.

[9] M. Hascoët and P. Dragicevic, "Interactive graph matching and visual comparison of graphs and clustered graphs," in *Proceedings of AVI2012*. ACM, May 2012.

[10] TreeTagger: a language independent part-of-speech tagger. [Online]. Available: http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/

[11] MeCab: Yet another part-of-speech and morphological analyzer. [Online]. Available: http://mecab.sourceforge.net

[12] T. Utsuro, Translation knowledge acquisition from cross-lingually relevant news articles, in *Proceedings of the 2nd China-Japan Natural Language Processing Promotion Conference*, 2002, pp. 123–134.

[13] OECD, *Assessing Scientific, Reading and Mathematical Literacy, A Framework for PISA2006*. OECD, 2006. http://www.oecd.org/france/PISA-2012-results-france.pdf and http://www.oecd.org/unitedkingdom/PISA-2012-results-UK.pdf

[14] NIER, *Assessing Scientific, Reading and Mathematical Literacy*. Gyosei, 2007.

[15] H. Nakasaki, M. Kawaba, D. Yokomoto, T. Utsuro, and T. Fukuhara, "Japanese english blog distribution and cross-lingual blog analysis with multilingual wikipedia entries as fundamental knowledge source," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 25, no. 5, pp. 613–622, 2010.

[16] Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.* 33, 1 (March 2001), 31-88. DOI=10.1145/375360.375365 http://doi.acm.org.gate6.inist.fr/10.1145/375360.375365

[17] Noack, A.: Energy models for graph clustering. J. Graph Algorithms Appl. 11(2), 453–480 (2007)