Deep Investigation of Cross-Language Plagiarism Detection Methods

Compilatio.net

IT SOLUTION FOR PLAGIARISM DETECTION

UNIVERSITÉ Grenoble Alpes

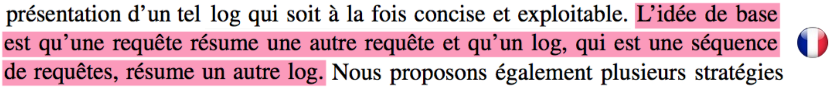
LIG

Authors

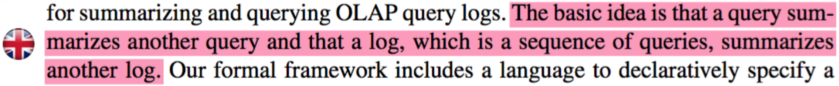
Jérémy Ferrero **Laurent Besacier** **Didier Schwab** **Frédéric Agnès**

What is Cross-Language Plagiarism Detection?

Cross-Language Plagiarism is a plagiarism by translation, *i.e.* a text has been plagiarized while being translated (manually or automatically).



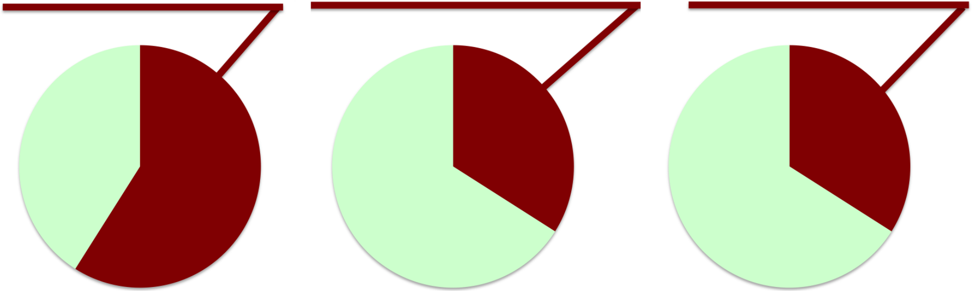
présentation d'un tel log qui soit à la fois concise et exploitable. L'idée de base est qu'une requête résume une autre requête et qu'un log, qui est une séquence de requêtes, résume un autre log. Nous proposons également plusieurs stratégies



for summarizing and querying OLAP query logs. The basic idea is that a query summarizes another query and that a log, which is a sequence of queries, summarizes another log. Our formal framework includes a language to declaratively specify a

From a text in a language L, we must find similar passage(s) in other text(s) from a set of candidate texts in language L' (cross-language textual similarity).

Why is it so important?



59% of high school students admitted cheating

34% doing it more than two times

1/3 admitted that they used the Internet to plagiarize

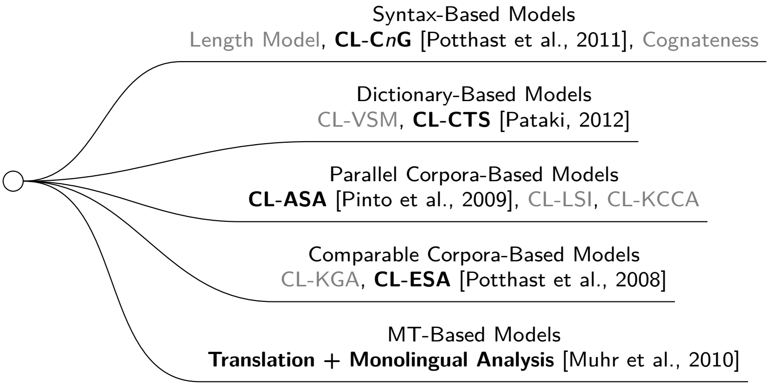
Sources:

- McCabe, D. (2010). Students' cheating takes a high-tech turn. In Rutgers Business School.
- Josephson Institute. (2011). What would honest Abe Lincoln say?

Research Questions

- How do the state-of-the-art methods behave according to the characteristics of the compared texts?
- Are the methods depend on the characteristics of the compared texts? And if so, which characteristics?
- Are the state-of-the-art methods complementary?

State-of-the-Art Methods



Syntax-Based Models

Length Model, **CL-C*n*G** [Potthast et al., 2011], Cognateness

Dictionary-Based Models

CL-VSM, **CL-CTS** [Pataki, 2012]

Parallel Corpora-Based Models

CL-ASA [Pinto et al., 2009], CL-LSI, CL-KCCA

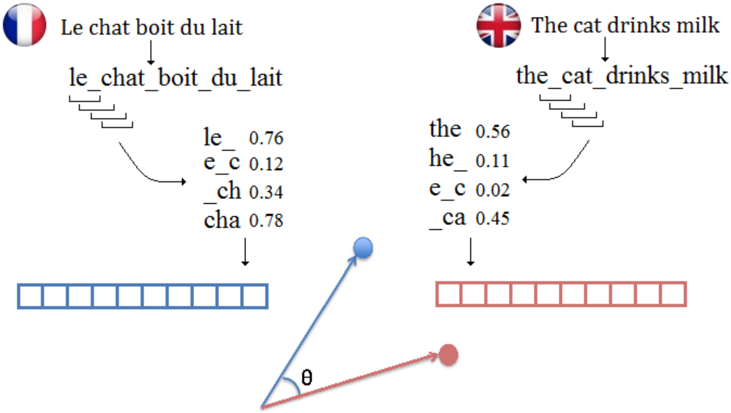
Comparable Corpora-Based Models

CL-KGA, **CL-ESA** [Potthast et al., 2008]

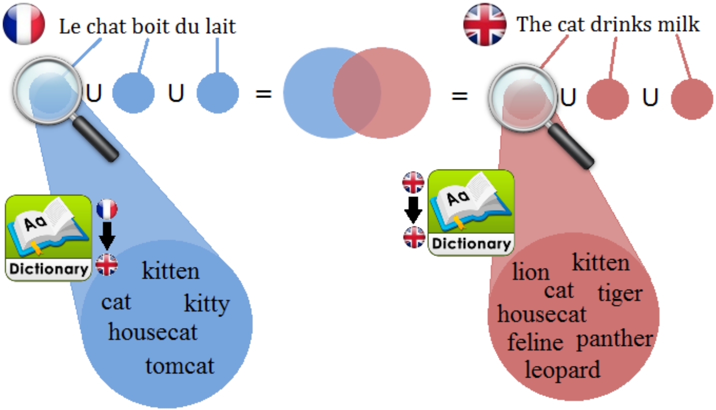
MT-Based Models

Translation + **Monolingual Analysis** [Muhr et al., 2010]

CL-C3G [Potthast et al., 2011]

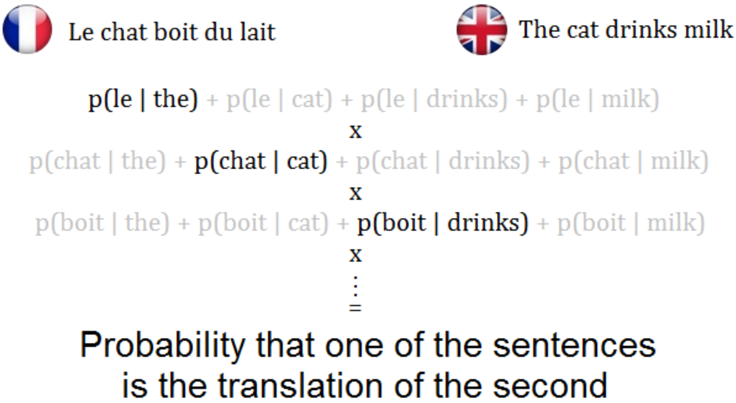


CL-CTS [Pataki, 2012]



We use DBNary [Sérasset, 2015] as linked lexical resource.

CL-ASA [Pinto et al., 2009]



Le chat boit du lait

The cat drinks milk

$$p(\text{le} \mid \text{the}) + p(\text{le} \mid \text{cat}) + p(\text{le} \mid \text{drinks}) + p(\text{le} \mid \text{milk})$$

$$\times$$

$$p(\text{chat} \mid \text{the}) + p(\text{chat} \mid \text{cat}) + p(\text{chat} \mid \text{drinks}) + p(\text{chat} \mid \text{milk})$$

$$\times$$

$$p(\text{boit} \mid \text{the}) + p(\text{boit} \mid \text{cat}) + p(\text{boit} \mid \text{drinks}) + p(\text{boit} \mid \text{milk})$$

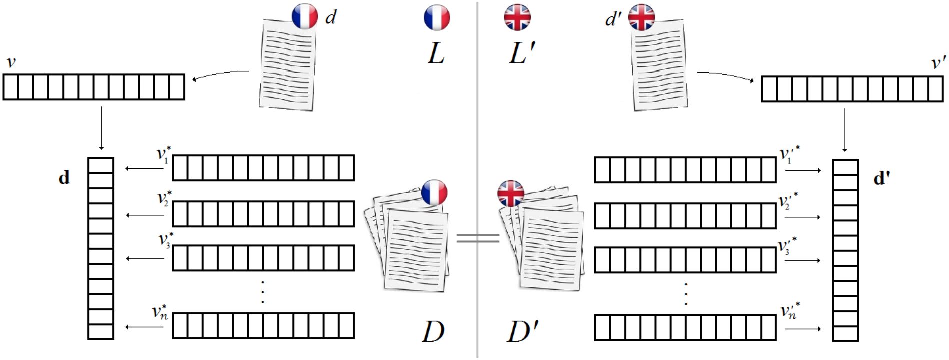
$$\times$$

$$\vdots$$

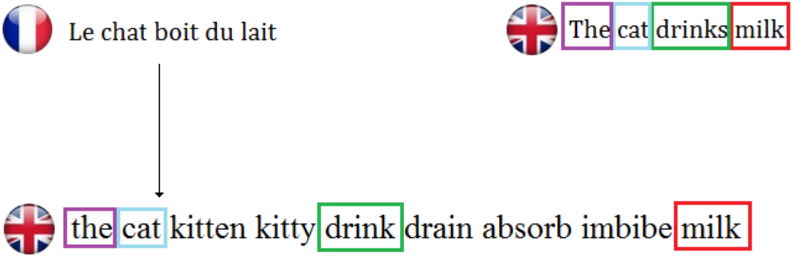
$$=$$

Probability that one of the sentences is the translation of the second

CL-ESA [Potthast et al., 2008]



T+MA [Muhr et al., 2010]



Le chat boit du lait

The cat drinks milk

the cat kitten kitty drink drain absorb imbibe milk

Evaluation Dataset [Ferrero et al., 2016][1]

- **French**, **English** and **Spanish**;
- **Parallel** and **comparable** (mix of Wikipedia, conference papers, product reviews, Europarl and JRC);
- Different granularities: **document** level, **sentence** level and **chunk** level;
- **Human** and **machine translated** texts;
- **Obfuscated** (to make the similarity detection more complicated) and **without added noise**;
- Written and translated by **multiple types of authors**;
- Cover **various fields**.

[1] A Multilingual, Multi-style and Multi-granularity Dataset for Cross-language Textual Similarity Detection. In Proceedings of LREC 2016. `https://github.com/FerreroJeremy/Cross-Language-Dataset`

Fist experiment: Evaluation Protocol

- We compared each textual unit to its corresponding unit in another language and to 999 other units randomly selected;
- We threshold the obtained distance matrix to find the threshold giving the best F_1 score;
- We repeat these two steps 10 times, leading to a 10 folds validation;
- The final value are the average of the 10 F_1 score.

Results: Across Language Pairs

Chunk level

Methods	EN→FR	FR→EN	EN→ES	ES→EN	ES→FR	FR→ES
CL-C3G	**0.5071**	**0.5071**	**0.4375**	**0.4375**	**0.4795**	**0.4795**
CL-CTS	0.4250	04116	0.3780	0.3881	0.4203	0.4169
CL-ASA	0.4738	0.4252	0.4083	0.3941	0.3736	0.3540
CL-ESA	0.1499	0.1499	0.1476	0.1476	0.1520	0.1520
T+MA	0.3730	0.3634	0.3177	0.3279	0.3158	0.3140

Sentence level

Methods	EN→FR	FR→EN	EN→ES	ES→EN	ES→FR	FR→ES
CL-C3G	**0.4931**	**0.4931**	**0.3819**	**0.3819**	0.4577	**0.4577**
CL-CTS	0.4734	0.4633	0.3171	0.3204	**0.4645**	0.4575
CL-ASA	0.3576	0.3523	0.2694	0.2531	0.3098	0.2843
CL-ESA	0.1430	0.1430	0.1337	0.1337	0.1383	0.1383
T+MA	0.3760	0.3692	0.3505	0.3526	0.3673	0.3525

Table 1: Overall F_1 score over all sub-corpora of the state-of-the-art methods for each language pair (EN: English; FR: French; ES: Spanish).

Results: Across Language Pairs

EN↔FR **EN↔ES**	**ES↔FR**
CL-C3G	CL-C3G
CL-ASA	CL-CTS
CL-CTS	CL-ASA

(a) Chunk granularity

EN↔FR **FR→ES**	**EN↔ES**	**ES→FR**
CL-C3G	CL-C3G	CL-CTS
CL-CTS	T+MA	CL-C3G
T+MA	CL-CTS	T+MA

(b) Sentence granularity

Table 2: Top 3 methods by source and target language.

Results: Across Language Pairs

Strong correlation between languages!

Chunk level							
EN→FR	**FR→EN**	**EN→ES**	**ES→EN**	**ES→FR**	**FR→ES**	**Overall**	**Lang. Pair**
1.000	0.991	0.998	0.995	0.957	0.940	0.980	**EN→FR**
	1.000	0.990	0.994	0.980	0.971	0.987	**FR→EN**
		1.000	0.996	0.967	0.949	0.983	**EN→ES**
			1.000	0.978	0.965	0.988	**ES→EN**
				1.000	0.998	0.980	**ES→FR**
					1.000	0.970	**FR→ES**

Sentence level							
EN→FR	**FR→EN**	**EN→ES**	**ES→EN**	**ES→FR**	**FR→ES**	**Overall**	**Lang. Pair**
1.000	1.000	0.929	0.922	0.991	0.982	0.971	**EN→FR**
	1.000	0.931	0.924	0.989	0.981	0.971	**FR→EN**
		1.000	0.997	0.925	0.913	0.949	**EN→ES**
			1.000	0.928	0.922	0.949	**ES→EN**
				1.000	0.997	0.971	**ES→FR**
					1.000	0.966	**FR→ES**

Table 3: Pearson correlations of the overall F_1 score over all sub-corpora of all methods between the different language pairs (EN: English; FR: French; ES: Spanish).

Results: Across Language Pairs

Strong correlation between granularities!

Lang. Pair	**Correlation**
EN→FR	0.907
FR→EN	0.946
EN→ES	0.833
ES→EN	0.838
ES→FR	0.932
FR→ES	0.939

Table 4: Pearson correlations of the results of all methods on all sub-corpora, between the chunk and the sentence granularity, by language pair (EN: English; FR: French; ES: Spanish) (calculated from Table 1).

Results: Across Language Pairs

Strong correlation between granularities!

Methods	Correlation
CL-C3G	0.996
CL-CTS	0.970
CL-ASA	0.649
CL-ESA	0.515
T+MA	0.780

Table 5: Pearson correlations of the results on all sub-corpora on all language pairs, between the chunk and the sentence granularity, by methods (calculated from Table 1).

Results: Detailed Analysis for English-French

Chunk level						
Methods	**Wikipedia (%)**	**TALN (%)**	**JRC (%)**	**APR (%)**	**Europarl (%)**	**Overall (%)**
CL-C3G	62.91 ±0.815	40.90 ±0.500	36.63 ±0.826	80.30 ±0.703	53.29 ±0.583	50.71 ±0.655
CL-CTS	58.00 ±0.519	33.71 ±0.382	29.87 ±0.815	67.51 ±1.050	44.95 ±1.157	42.50 ±1.053
CL-ASA	23.33 ±0.724	23.39 ±0.432	33.14 ±0.936	26.49 ±1.205	55.50 ±0.681	47.38 ±0.781
CL-ESA	64.89 ±0.664	23.78 ±0.613	14.03 ±0.997	23.14 ±0.777	14.19 ±0.590	14.99 ±0.709
T+MA	58.22 ±0.756	39.13 ±0.551	28.61 ±0.597	73.14 ±0.666	36.95 ±1.502	37.30 ±1.200

Sentence level						
Methods	**Wikipedia (%)**	**TALN (%)**	**JRC (%)**	**APR (%)**	**Europarl (%)**	**Overall (%)**
CL-C3G	48.25 ±0.349	48.08 ±0.538	36.68 ±0.693	61.10 ±0.581	52.72 ±0.866	49.31 ±0.798
CL-CTS	46.68 ±0.437	38.67 ±0.552	28.21 ±0.612	50.82 ±1.034	53.21 ±0.601	47.34 ±0.632
CL-ASA	27.63 ±0.330	27.25 ±0.341	35.17 ±0.644	25.53 ±0.795	36.55 ±1.139	35.76 ±0.978
CL-ESA	51.14 ±0.875	14.25 ±0.334	14.44 ±0.341	13.93 ±0.714	13.91 ±0.618	14.30 ±0.551
T+MA	50.57 ±0.888	37.79 ±0.364	32.36 ±0.369	61.94 ±0.756	37.92 ±0.552	37.60 ±0.518

Table 6: Average F_1 scores and confidence intervals of methods applied on EN→FR sub-corpora at chunk and sentence level – 10 folds validation.

Second Experiment: Evaluation Protocol

- We compare 1000 English textual units to their corresponding unit in French, and to one other (not relevant) French unit;
- Each unit must strictly leads to one match and one mismatch (= 1000 matches and 1000 mismatches);
- We repeat these two steps 10 times, leading to a 10 folds validation.

Complementarity?

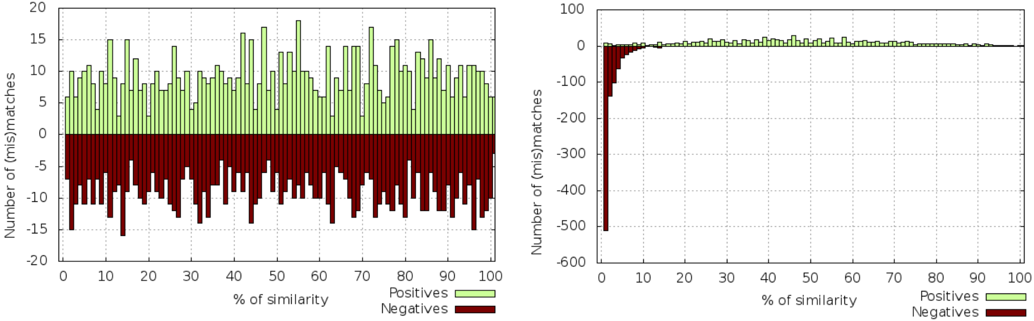


Figure 1: Distribution histograms of *Random Baseline* (left) and *CL-C3G* (right) for 1000 positives (lightgreen) and 1000 negatives (darkred) (mis)matches.

Complementarity?

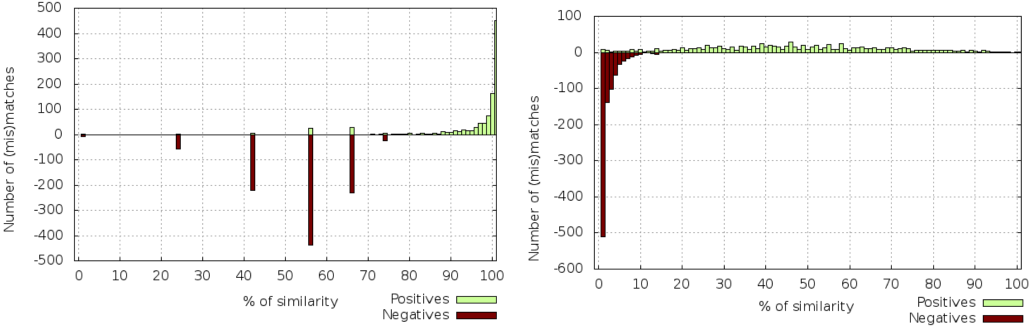
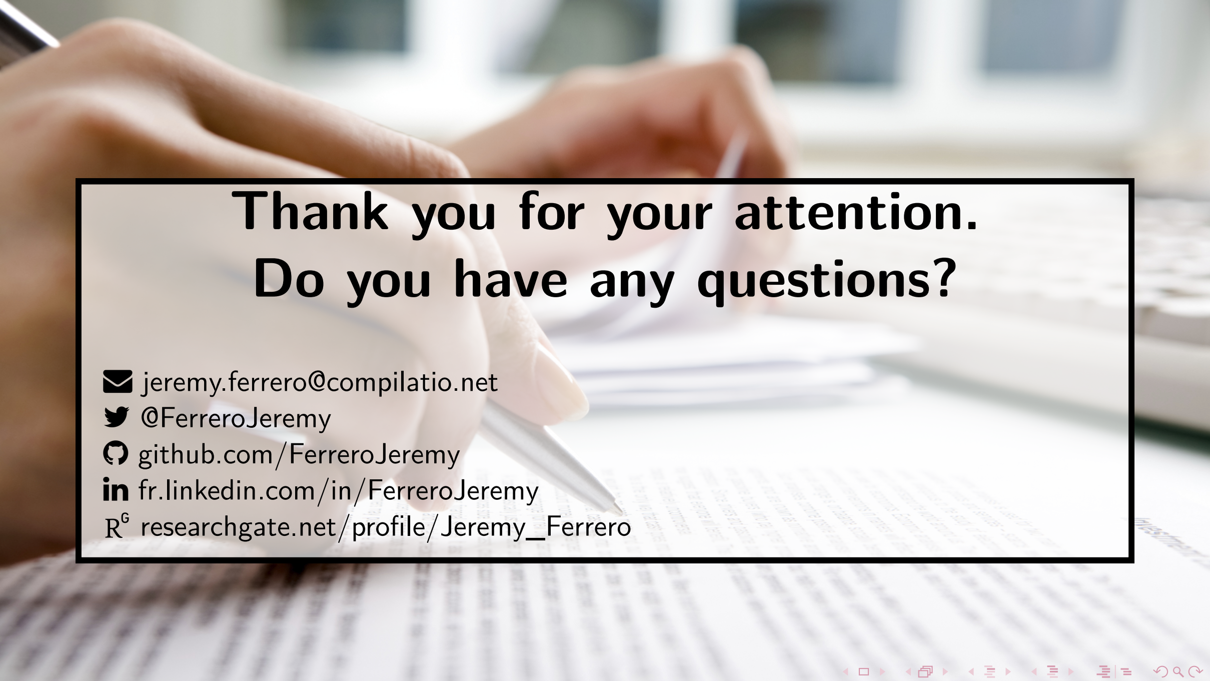


Figure 2: Distribution histograms of *CL-ASA* (left) and *CL-C3G* (right) for 1000 positives (lightgreen) and 1000 negatives (darkred) (mis)matches.

Conclusion

- Results show a common behavior of methods across different language pairs;
- Strong correlations across languages, sizes and types of texts;
- Methods behave differently in clustering, even if they seem similar in performance ⇒ combination or fusion?

I invit you to come see my poster this afternoon at SemEval workshop to verify that ;)



Thank you for your attention.
Do you have any questions?

✉ jeremy.ferrero@compilatio.net
@FerreroJeremy
github.com/FerreroJeremy
in fr.linkedin.com/in/FerreroJeremy
R^G researchgate.net/profile/Jeremy_Ferrero

References I

- Ferrero, J., Agnès, F., Besacier, L., and Schwab, D. (2016).
  A Multilingual, Multi-style and Multi-granularity Dataset for Cross-language Textual Similarity Detection.
  In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4162–4169, Portoroz, Slovenia. European Language Resources Association (ELRA).

- Muhr, M., Kern, R., Zechner, M., and Granitzer, M. (2010).
  External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System - Lab Report for PAN at CLEF 2010.
  In Braschler, M., Harman, D., and Pianta, E., editors, *CLEF Notebook*, Padua, Italy.

References II

Pataki, M. (2012).
A New Approach for Searching Translated Plagiarism.
In *Proceedings of the 5th International Plagiarism Conference*, pages 49–64, Newcastle, UK.

Pinto, D., Civera, J., Juan, A., Rosso, P., and Barrón-Cedeño, A. (2009).
A Statistical Approach to Crosslingual Natural Language Tasks.
In *CEUR Workshop Proceedings*, volume 64 of *Journal of Algorithms*, pages 51–60.

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011).
Cross-Language Plagiarism Detection.
In *Language Resources and Evaluation*, volume 45, pages 45–62.

References III

Potthast, M., Stein, B., and Anderka, M. (2008).
A Wikipedia-Based Multilingual Retrieval Model.
In *30th European Conference on IR Research (ECIR'08)*, volume 4956 of *LNCS of Lecture Notes in Computer Science*, pages 522–530, Glasgow, Scotland. Springer.

Sérasset, G. (2015).
DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF.
In *Semantic Web Journal (special issue on Multilingual Linked Open Data)*, volume 6, pages 355–361.