The Varitext platform and the Corpus des variétés nationales du français (CoVaNa-FR) as resources for the study of French from a pluricentric perspective

Sascha Diwersy
Institute of Romance Languages / Centre de Recherche Interdisciplinaire sur la France et la Francophonie (CIFRA) University of Cologne
sascha.diwersy@uni-koeln.de

Abstract

This paper reports on the francophone corpus archive Corpus des variétés nationales du français (CoVaNa-FR) and the lexico-statistical platform Varitext. It outlines the design and data format of the samples as well as presenting various usage scenarios related to the applications featured by the platform’s toolbox.

1 Introduction

This contribution presents the francophone corpus archive Corpus des variétés nationales du français (CoVaNa-FR) and its hosting platform Varitext.

The paper is structured as follows. Section 2 will outline the rationale behind the corpus archive, its composition and its data format. In section 3, we will then introduce the toolbox implemented by the Varitext platform, by illustrating some of its functionalities and giving brief sketches of corresponding usage scenarios. Section 4 provides a brief summary and discusses possible directions for the future development of the resources presented in this paper.

2 The CoVaNa-FR corpus archive

2.1 Rationale and composition of the CoVaNa-FR

The creation of the Corpus des variétés nationales du français (CoVaNa-FR) is motivated by the aim of offering a large-scale resource to researchers working on the French language from a pluricentric perspective. It is thus primarily designed to provide methodological support for investigations in the French tradition of ‘lexicologie différencielle’ (‘variationist differential lexicography’) focusing on elements of endonormative differentiation, i.e. the emergence of regionally specific norms compared to a supposed metropolitan standard variety of French (for studies on various francophone regions, see Rézeau 2007, Thibault 2008; for studies especially focusing on Subsaharan Africa and the Maghreb, cf. Queffélec 1997, Laftae 2002, Naffati and Queffélec 2004, Nzesse 2009, to mention just a few examples of a sizable body of literature). Alongside the lexico-statistical toolbox implemented by the Varitext platform (cf. Section 3 below), the design of the CoVaNa-FR goes beyond the rather conventional lexicological rationale of the lexicological framework just mentioned and can be seen as a contribution to meeting the desideratum, voiced by Stein (2003:14f), of carrying out large-scale investigations on Francophone varieties using contemporary corpus linguistic methods. In this regard, the CoVaNa-FR differs from existing French corpora such as Frantext (cf. ATILF-CNRS), Québétext (cf. Trésor de la langue française au Québec) and Suistext (cf. Trésor des Vocabulaires francophones Neuchâtel) in offering broad regional coverage (bundling samples from Africa, Europe and North America), a wider range of query functionalities and free access (large parts of Frantext not being accessible free of charge and Suistext only being available locally at its hosting institution, cf. Thibault 2007:480). Apart from corpus linguistic uses, the CoVaNa-FR could also be a valuable resource for research on the automatic classification of language varieties, which has recently aroused considerable interest in the field of NLP (for relevant contributions see, amongst others, Ranaivo-Malancon 2006, Ljubešić et al. 2007, Tiedemann and Ljubešić 2012,

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: http://creativecommons.org/licenses/by/4.0/
Trieschnigg et al. 2012, Zampieri and Gebre 2012, Tan et al. 2014). It should be noted, though, that in accordance with copyright restrictions, the CoVaNa-FR is not directly available for download and can only be consulted via the GUI of the password-protected Varitext platform.

Due to its focus on endonormative differentiation, the CoVaNa-FR is less balanced with respect to genre than similar corpora for other languages such as the International Corpus of English (ICE, cf. Greenbaum 1996), the Corpus of Contemporary American English (COCA, cf. Davies 2009), the Corpus de Referencia del Español Actual (CREA, cf. Real Academia Española), the Corpus del Español (cf. Davies 2002) or the Corpus do Português (cf. Davies 2014).1 The initial version of the CoVaNa-FR, accessible on the Varitext platform, is made up of journalistic texts published by national newspapers in different Francophone countries in Africa, Europe and North America. The choice of national newspapers as primary sources is based on the assumption made by Glessgen (2007:97) that these are particularly representative of contemporary standard varieties (“les grands journaux [...] reflètent assez bien les variétés standard actuelles”). Work is also underway on the extension of the CoVaNa-FR, such that future versions will include a subcorpus of fiction and academic texts. In its present state, the CoVaNa-FR is divided into 11 samples collected across a span of at least two years and categorized by regional parameters as listed in Table 1.

| Sample code | Country           | Sources                                                                 | Number of word tokens² |
|-------------|-------------------|-------------------------------------------------------------------------|------------------------|
| DZA         | Algeria           | El Watan, La Tribune d’Alger                                           | 45,600,000             |
| CAM         | Cameroon          | Cameroon Tribune, La Nouvelle Expression, Mutations                    | 46,500,000             |
| CAN         | Canada (Québec)   | Le Devoir, Le Soleil                                                  | 53,500,000             |
| COD         | Congo (D.R.C.)    | Le Potentiel                                                           | 27,300,000             |
| FRA         | France            | Le Figaro, Le Monde                                                   | 53,300,000             |
| CIV         | Ivory Coast       | Fraternité Matin, Notre Voie                                          | 18,800,000             |
| MLI         | Mali              | Aurore, L’Essor, L’Indépendant                                         | 25,100,000             |
| MAR         | Morocco           | Aujourd’hui le Maroc, Le Matin du Sahara                               | 43,600,000             |
| SEN         | Senegal           | Le Soleil, Wal Fadjiri                                                | 27,100,000             |
| CHE         | Switzerland       | Le Temps, La Tribune de Genève                                         | 28,000,000             |
| TUN         | Tunisia           | Le Presse, Le Quotidien, Le Temps                                     | 50,900,000             |
| Total       |                   |                                                                        | 419,700,000            |

Tab. 1: Composition of the CoVaNa-FR (on-line version accessible via the Varitext platform).

The compilation of the overall corpus archive outlined in Table 1 has been carried out according to the requirement that each country be represented by a sample comprising at least two newspapers with articles from the same (or similar) two years. It should be noted, though, that some samples do not fully meet these guidelines, as is the case with the corpora representing Algeria and Canada (containing two newspapers from single and different years) or the sample representing the Democratic Republic of Congo (containing three years of only one newspaper).

2.2 Processing format of the CoVaNa-FR

All documents in the CoVaNa-FR corpus are formatted in eXtensible Markup Language (XML) with the structural units (i) subcorpus, (ii) text, (iii) paragraph, and (iv) sentence. The texts are annotated with (i) part-of-speech (PoS) tags, (ii) lemmas and (iii) dependency-parses using the commercially licensed Connexor annotation tool (Tapanainen and Järvinen 1997). The corpus files are in standard CWB input format (cf. Evert and Hardie 2011:5f) with XML tags and each token record (one surface form + associated TAB-delimited token-level annotations) appearing on separate lines.

The set of XML tagged structural units is specified by the DTD given in Figure 1. Note that the top level <corpus>...</corpus> element defines one country related sample and that each subcorpus corresponds to a one year newspaper volume. The element attributes which are provided inside the query

---

1See the projects’ web sites at http://ice-corpora.net/ICE/INDEX.HTM, http://corpus.byu.edu/coca/, http://corpus.rae.es/creanet.html, http://www.corpusdelespanol.org and http://www.corpusdoportugues.org respectively.

2Numbers are rounded down to the nearest 100,000.
platform as metadata categories for corpus partitioning or the description of concordance extracts are highlighted in boldface.

<!DOCTYPE varcorpus [  
<!-- country related sample -->  
<!ELEMENT corpus (subcorpus)+>  
<!-- one year newspaper volume -->  
<!ELEMENT subcorpus (text)+>  
<!-- newspaper article -->  
<!ELEMENT text (p)+>  
<!-- paragraph -->  
<!ELEMENT p (s)+>  
<!-- sentence -->  
<!ELEMENT s (#PCDATA)>  
<!ATTLIST corpus id CDATA #REQUIRED  
name CDATA #REQUIRED  
code CDATA #REQUIRED  
geocode CDATA #REQUIRED  
geoname CDATA #REQUIRED  
]>

<!ATTLIST subcorpus id CDATA #REQUIRED  
name CDATA #REQUIRED  
code CDATA #REQUIRED  
source CDATA #REQUIRED  
year CDATA #REQUIRED  
]>

<!ATTLIST text id CDATA #REQUIRED  
title CDATA #REQUIRED  
author CDATA #REQUIRED  
date CDATA #REQUIRED  
section CDATA #REQUIRED  
]>

<!ATTLIST p id CDATA #REQUIRED  
type CDATA #IMPLIED  
]>

<!ATTLIST s id CDATA #REQUIRED>

]> Fig. 1: DTD specifying the structural elements of the country-related samples in the CoVaNa-FR corpus archive.

As for the token rows, their core structure is basically defined according to the so-called CoNLL format, introduced on the occasion of the correspondent 2007 shared task on dependency parsing (cf. Nivre et al. 2007:916). For rather technical reasons, this structure has been extended by a number of fields whose purpose is to optimize the processing of queries exploring the dependency relations annotated in the corpus. The fields in question are marked by an asterisk in the following table, which outlines the overall structure of the token records:

| Field name | Description |
|------------|-------------|
| id         | sentence internal numerical token identifier (counter starting at 1 for each sentence) |
| word       | surface form or punctuation sign |
| lemma      | lemma corresponding to the surface form |
| cpos       | coarse grained part of speech (PoS) |
| pos        | fine grained PoS + morphological features |
| headid     | token identifier of the syntactic head |
| headoffset | distance between syntactic head and token |
| deprel     | syntactic function of the token in the dependency relation to its head |
| headword   | surface form of the syntactic head |
| headlemma  | lemma of the syntactic head |
| headcpos   | coarse grained PoS of the syntactic head |
| headpos    | fine grained PoS + morphological features of the syntactic head |
| pmarkword  | surface form of the function word (adposition or conjunction) dependent on the token |
| pmarklemma | lemma of the function word dependent on the token |
| pmarkcpos  | PoS of the function word dependent on the token |

Tab. 2: Structure of the token records contained by the corpus files.
The 11 country specific samples making up the present online version of the CoVaNa-FR (see Table 1 above) have been encoded by means of the IMS Open Corpus Workbench (CWB, cf. Evert and Hardie 2011; see also the project’s web site http://cwb.sourceforge.net/), the total size of the corresponding index files summing up to 58.4 GB of disk space. The components of CWB are integrated as main query processing tools in the Varitext platform, which will be described in more detail in the following section.

3 The Varitext platform

3.1 Design and GUI

Varitext is a web-based platform (cf. http://syrah.uni-koeln.de/varitext/ and http://extranet-ldi.univ-paris13.fr/varitext/) providing free-of-charge access to the CoVaNa-FR corpus archive presented in section 2. As is indicated by its name, it is open to host corpora for other languages compiled according to the same rationale of large-scale variationist research in a pluricentric perspective. Work has already been completed on the prototype of a hispanophone corpus archive, which will be released via Varitext in the near future. There are also plans to compile similar resources for Portuguese, Russian and Arabic.

The toolbox implemented by the Varitext platform is built upon three major software components: CWB for query processing, the UCS toolkit version 0.6 (cf. Evert 2005, the software being available at http://www.collocations.de/software.html) for cooccurrence analysis and R (R Core Team 2014) for statistical computing and plotting.

The platform’s user interface allows fairly complex queries in terms of subsampling and the formulation of search expressions. Using the menu options relating to the available metadata categories (such as country code, newspaper volume or thematic section), it is possible to create subcorpora and partitions with different degrees of granularity, as is shown by Fig. 2:

![Fig. 2: Using menu options to build a partition defined by country on the basis of a subcorpus comprising the samples representing Cameroon and the Ivory Coast and filtered by the thematic section ‘Sports’](image)

As for the formulation of query expressions, the interface integrates a sub-menu to set up search constraints flexibly by combining several token properties (such as lemma, PoS or syntactic function; see the data model outlined in table 2 above) and / or assembling sequences of various length (see Figure 3).

---

The annotation model of Connexor treats adpositions and conjunctions as markers dependent on content words (verbs, nouns, adjectives, adverbs).
In its present state, the Varitext platform features as its standard applications a KWIC concordancer and a set of tools for frequency computing, key word analysis and collocation processing, the latter of which will be outlined in some detail below. Future releases of the platform will also include advanced functionalities of statistical computing and plotting that are currently under development and testing and which will be briefly sketched at the end of this section.

3.2 Usage Scenario: Sample Specific Frequencies and Lexical Differences

3.2.1 chaussure vs. soulier

One of the platform’s standard applications besides KWIC concordancing is the computation of sample specific frequencies and key word analysis. In a corpus-based perspective, these methods can be used for instance as diagnostics to test the results of ‘differential’ lexicology. Similar to Thibault’s (2007) study on some lexical specificities of Canadian (Quebec), Swiss and metropolitan standard French, it would be possible to analyze geographical lexical variants in terms of their frequency distribution. An example also mentioned by Thibault (2007:468-475) is provided by the nouns chaussure and soulier (“shoe”), with soulier being regarded as regional variant especially of Canadian French (cf. the reference dictionary Le Petit Robert (Rey-Debove and Rey 2006) s.v. SOULIER). A key word analysis based on the samples representing Canada/Quebec (geographical code: CAN), France (FRA) and Switzerland (CHE) yields the log-likelihood ratio (LL) scores given by the following bar plots in Fig. 4 (for the use of the log likelihood ratio in key word analysis see Rayson 2003). The computation has been carried out on a 2x2 basis, with one sample as the main corpus and the combination of the remaining two as the reference corpus.

Fig. 4: LL scores for the nouns chaussure and soulier in the samples representing Canada/Quebec, Switzerland and France
These figures indicate that there are clear-cut distributional divergences, with the two nouns being respectively under- and overrepresented in the samples related to Quebec and France. This seems to suggest that *soulier* is still part of the French standard as it evolves in Quebec, or at least in its national newspapers, which qualifies to some extent the findings of Thibault (2007:474), according to which Quebec newspaper language is moving towards greater conformity with French metropolitan usage in the case of *chaussure* and *soulier*. It should be noted that Thibault only considers the relative frequencies of the two items within each national sample. Applying this approach to our corpus data would provide no more than a confirmation of Thibault’s findings. In light of the aforementioned key word analysis, though, there is sufficient evidence to conclude that, in Quebec French, the relationship between the two variants is rather more complex and should be subjected to a more detailed analysis in terms of collocational distribution. One promising approach in this respect would be Hoey’s (2005) lexical priming theory.

### 3.2.2 Quebec Specific Lexical Items

At this point, it is worth noting that, although major national newspapers might reflect trends of standard varieties quite faithfully (see our reference to Glessgen 2007 in section 2), the data obtained from these sources should be handled with some caution (cf. also Thibault 2007:474). This is of particular importance if we adopt a corpus-driven approach, which involves identifying the most characteristic features in a sample by means of statistical techniques such as key word analysis.

This may be illustrated with the results of a key word analysis contrasting the Quebec sub-corpus as a whole with the sample representing France.

| Lemma       | Frequency CAN | Frequency FRA | Rel. Freq. CAN | Rel. Freq. FRA | LL score | Rank |
|-------------|---------------|---------------|---------------|---------------|----------|------|
| Québec      | 93269         | 828           | 1740.4        | 15.53         | 120592.82| 1    |
| Montréal    | 44257         | 472           | 825.83        | 8.85          | 56578.51 | 2    |
| Canada      | 43612         | 1808          | 813.8         | 33.9          | 47579.89 | 3    |
| :           | :             | :             | :             | :             | :        |      |
| chum        | 1191          | 4             | 22.22         | 0.08          | 1597.32  | 243  |
| :           | :             | :             | :             | :             | :        |      |
| magasiner   | 183           | 1             | 3.41          | 0.02          | 241.78   | 1987 |
| :           | :             | :             | :             | :             | :        |      |
| placoter    | 18            | 0             | 0.34          | 0             | 24.87    | 10744|
| :           | :             | :             | :             | :             | :        |      |
| paqueter    | 13            | 0             | 0.24          | 0             | 17.96    | 13473|
| :           | :             | :             | :             | :             | :        |      |

Tab. 3: Words specific to the Quebec sub-corpus in contrast with the sample representing France.

The data given in Table 3 show that the most specific items are proper nouns closely related to socio-cultural context, whereas words which clearly qualify as Quebecisms, such as *chum* ("friend, pal"), *magasiner* ("to go shopping"), *placoter* ("to chat"; cf. Poirier 1995:32) or *paqueter* ("to pack"; cf. Poirier ibid) only come at lower ranks, their log-likelihood scores being nonetheless highly significant.

### 3.3 Usage Scenario: Lexical Cooccurrences and Collocational Variation

The second main application provided by the platform’s toolbox is collocation analysis. We will illustrate this functionality by considering the example of the causative support verb *occasionner* ("to occasion sth") and the semantic associations instantiated by its most significant collocates within each of the

---

4Figures are given in terms of token per million.
samples making up the CoVaNa-Fr corpus archive.

The following cross table which is based on the lexicogram (defined as list of collocates specified by association scores; see Tournier 1987) computed for occasionner displays some of the nouns in direct object position significantly collocating with this verb in terms of the log-likelihood ratio (the use of the latter as an association measure for collocation analysis having been proposed, amongst others, by Dunning 1993).

| Collocate | CAN | CHE | CIV | CMR | COD | DZA | FRA | MAR | MLI | SEN | TUN |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| accident  | -   | -   | -   | 67.8| -   | 65.4| -   | -   | 61.2| -   | -   |
| accroissement | - | - | - | 68.5| - | - | - | - | - | - | - |
| augmentation | - | - | - | 52.4| - | - | - | - | - | - | - |
| baisse    | -   | -   | -   | 41.7| -   | -   | -   | -   | 59.5| -   | -   |
| coût      | 90.3| -   | -   | -   | -   | -   | -   | -   | -   | -   | -   |
| dégât     | -   | 87.6| -   | 91.8| 268.5| 1059.3| 62.3| 255.7| 157.6| 208.5| 143.9|
| perte     | 298.8| 109.4| 267.8| 178.0| 208.8| 381.4| 64.9| 134.1| 492.9| 170.5| 129.7|
| problème  | 62.37| - | - | - | - | 23.1| - | - | - | - | 33.1|

Tab. 4: Significant direct object noun collocates of occasionner across all the samples contained by the CoVaNa-FR.

It is easy to see that the combinatorial profile of occasionner is essentially characterized by negative semantic prosody throughout all the samples under investigation (for the concept of semantic prosody, see Stubbs 1995 and Xiao and McEnery 2006). At the same time, however, it exhibits some degree of regional variation; in the case of the sub-corpus representing the Congo (COD), for example, there is an additional semantic feature in evidence which may be described as INTENSITY (cf. the collocates accroissement [“increase, growth”], augmentation [“increase, rise”] and baisse [“decrease, fall”]).

A similar statement can be made with regard to the significant noun collocates of causer (“to cause”), although in this case it is the Quebec sample which adds more neutral marked elements (surprise [“surprise”]) to the overall picture. We illustrate this by a means of a plot generated by a correspondence analysis (CA, see Lebart et al. 1998:47ff) performed on the sample specific lexicograms comprising the direct object nouns significantly associated with the verb in question (further examples of using CA to explore the CoVaNa-FR are given by Diwersy and Loiseau forthcoming):

![CA factor map](image)

Fig. 5: Plot generated by a CA performed on the country specific lexicograms of causer.

---

5 Sample name as translated to their corresponding ISO 3166-1 alpha-3 country codes (see the UN Statistic Division’s page at http://unstats.un.org/unsd/methods/m49/m49alpha.htm).

6 The collocates used for further processing have been selected according to a frequency threshold of 20 and an LL score threshold of 10.83.
The CA plot given in Fig. 5 highlights in its main (horizontal) dimension the contrast between the Quebec subcorpus and the remaining samples, this contrast being paralleled by the contrast between the noun *surprise* and other items such as *souci* (“worry”) and *dégât* (“damage”).

Correspondence analysis is a useful technique in providing a condensed view of divergences relating to samples and lexical items. It will be included in the next release of the Varitext platform.

### 4 Conclusion

As the examples in the preceding section have shown, there is considerable scope for using corpus-related techniques (beyond concordancing) to investigate geographical variation from a pluricentric perspective, but researchers must exercise caution when working on the diverse sets of data which can be obtained using the resources outlined in this paper. A major case in point is the composition of the corpus archive and its current restriction to journalistic texts, which may bring about phenomena related to the socio-cultural context rather than the linguistic one (although, from the point of view of media discourse analysis and communication studies, these thematic „side effects“ could be of quite some interest).

It should be obvious, then, that our present activities focus on diversifying the corpus resources, especially with regard to other written genres. At the same time, we are engaged in extending the overall text archive to include corpora for different languages, the rationale being to apply the methodological framework implemented by the Varitext platform to linguistic areas other than Francophonie.

This framework is itself undergoing considerable modifications which will lead to the integration of advanced statistical functionalities. At present, our main interest is to enhance the platform’s toolbox by implementing several exploratory multivariate techniques, which will be tested in experimental settings that, however, go beyond the narrow focus of this paper.

That said, the development of the corpus archive and of the platform is still in its infancy, and is set to evolve further in various ways and directions. At least, this is what should happen if the community makes good use of it.

### Acknowledgements

The author wishes to thank the reviewers for their valuable comments which helped to clarify the main points of the paper.

### References

ATILF-CNRS. *Base textuelle FRANTEXT*. ATILF-CNRS Nancy & Université de Lorraine. http://www.frantext.fr/.

Mark Davies. 2002. Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *SEPLN 2002* (Sociedad Española para el Procesamiento del Lenguaje Natural), 21–27.

Mark Davies. 2009. The 385+ Million Word Corpus of Contemporary American English (1990-2008+): Design, Architecture, and Linguistic Insights. *International Journal of Corpus Linguistics*, 14: 159–190.

Mark Davies. 2014. Creating and Using the Corpus do Português and the Frequency Dictionary of Portuguese. Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira (eds.): *Working with Portuguese Corpora*. London: Bloomsbury Publishing, 89–110.

Sascha Diwersy and Sylvain Loiseau. Forthcoming. La différenciation du français dans l’espace francophone: l’apport des statistiques lexicales. Kirsten A. Jeppesen Kragh, Jan Lindschouw and Rene Schøsler (eds.): *Les variations diasystématiques dans les langues romanes et leurs interdépendances*. Société de Linguistique Romane.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61–74.

---

7The given CA plot has been generated by means of the R package FactoMineR (cf. Husson et al. 2013).

8To be more precise, the main dimension (read from right to left) puts into contrast nouns opposed by the features (1) ‘neutral’ vs. ‘negative’ (affect) polarity (surprise vs. souci), (2) ‘physical’ vs. ‘material’ damage (lésion [“injury, lesion”] vs. dégât / préjudice [“damage”]) and (3) ‘non-lethal’ vs. ‘lethal’ impact (lésion vs. mort [“death”]).
Stefan Evert. 2005. Empirical research on association measures: The UCS toolkit. *Software demonstration at the Phraseology 2005 Conference*, Louvain-la-Neuve, Belgium. [Abstract available at http://purl.org/stefan.evert/PUB/Evert2005phraseology.pdf]

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, UK. [PDF version available for download at http://www.birmingham.ac.uk/documents/college-arts-law/corpus/conference-archives/2011/Paper-153.pdf]

Martin-Dietrich Glessgen. 2007. *Linguistique romane, domaine et méthode – Domaines et méthodes en linguistique française et romane*. Paris: Armand Colin.

Sidney Greenbaum (ed.). 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Michael Hoey. 2005. *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

François Husson, Julie Josse, Sebastien Lê and Jeremy Mazet. 2013. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. http://CRAN.R-project.org/package=FactoMineR.

Suzanne Lafage. 2002. *Le lexique français de Côté-d’Ivoire* (Appropriation et créativité). Nice: CNRS.

Ludovic Lebart, André Salem and Lisette Berry. 1998. *Exploring Textual Data*. Dordrecht: Springer.

Nikola Ljubešić, Nives Mikelić and Damir Boras. 2007. Language Identification: How to Distinguish Similar Languages? *Proceedings of the 29th International Conference on Information Technology Interfaces*, Zagreb, Croatia.

Habiba Naffati and Ambroise Queffélec. 2004. *Le français en Tunisie*. Nice: CNRS.

Claude Poirier. 1995. *Les variantes topolectales du lexique français: propositions de classement à partir d’exemples québécois*. Michel Francard and Danièle Latin (eds.): *Le régionalisme lexical*. Louvain-la-Neuve: De Boeck, 13–56.

Ambroise Queffélec. 1997. *Le français en Centrafrique: lexique et société*. Vanves: Editions Classiques d’Expression Française (EDICEF).

Paul Rayson. 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University. [PDF version available for download at http://ucrel.lancs.ac.uk/people/paul/publications/phd2003.pdf]

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. [http://www.R-project.org/]

Josette Rey-Debove and Alain Rey (eds.). 2006. *Le Nouveau Petit Robert: Dictionnaire alphabétique et analogique de la langue française*. Paris: Dictionnaires Le Robert.

Pierre Rézeau (ed.). 2007. *Richesse du français et géographie linguistique*, volume 1. Louvain-la-Neuve: de Boeck.

Achim Stein. 2003. Lexikalische Köökurrenz im afrikanischen Französisch. *Zeitschrift für französische Sprach- und Literaturwissenschaft*, 113: 1–17.

Michael Stubbs. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1): 23–55.
Liling Tan, Marcos Zampieri, Nikola Ljubešić and Jörg Tiedemann. 2014. Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. *Proceedings of the 7th Workshop on Building and Using Comparable Corpora: Building Resources for Machine Translation Research*, Reykjavik, Iceland.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, 64–74.

André Thibault. 2007. Banques de données textuelles, régionalismes de fréquence et régionalismes négatifs. *ACILPR XXIV*, volume 1, 467–480.

André Thibault (ed.). 2008. *Richesse du français et géographie linguistique*, volume 2. Louvain-la-Neuve: de Boeck.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. *Proceedings of COLING 2012*, Mumbai, India, 2619–2634.

Trésor de la langue française au Québec. *Base textuelle QUEBÉTEXT*. Université Laval, Département de Langues, linguistique et traduction. [http://www.tlfq.ulaval.ca/quebetext/](http://www.tlfq.ulaval.ca/quebetext/)

Trésor des Vocabulaires francophones Neuchâtel. *Base textuelle SUISTEXT*. Université de Neuchâtel, Centre de dialectologie et d’étude du français régional.

Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Franciska de Jong and Theo Meder. 2012. An exploration of language identification techniques for the Dutch Folktale Database. *Proceedings of LREC 2012*, Istanbul, Turkey.

Maurice Tournier. 1987. Cooccurrences autour de travail (1971-1976). *Mots*, 14: 89–123.

Richard Xiao and Tony McEnery. 2006. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied linguistics*, 27(1): 103–129.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic Identification of Language Varieties: The Case of Portuguese. *Proceedings of KONVENS 2012*, Vienna, Austria, 233–237.