ride. A review journal for digital editions and resources

published by the IDE

Varitext

Varitext, Sascha Diwersy, Peter Blumenthal, Salah Mejri (ed.), 2013. http://syrah.uni-koeln.de/varitext/ (Last Accessed: 15.08.2017). Reviewed by Julia Burkhardt (University of Leipzig), jburk (at) rz.uni-leipzig.de.

Abstract

This review discusses the web-based platform *Varitext* that up to now provides free-of-charge access to the *Corpus des variétés nationales du français (CoVaNa-FR)* and is designed to assimilate further linguistic corpora of other languages in the future. The aim of *Varitext* is to make available large-scale resources of so-called 'pluricentric' languages focussing on the national variations of the respective standard languages. At present French standard varieties in CoVaNa-FR are represented by texts of the francophone daily press. *Varitext* offers a working environment making it possible to analyse the corpus of standard varieties with primary regard to lexical and statistical aspects.

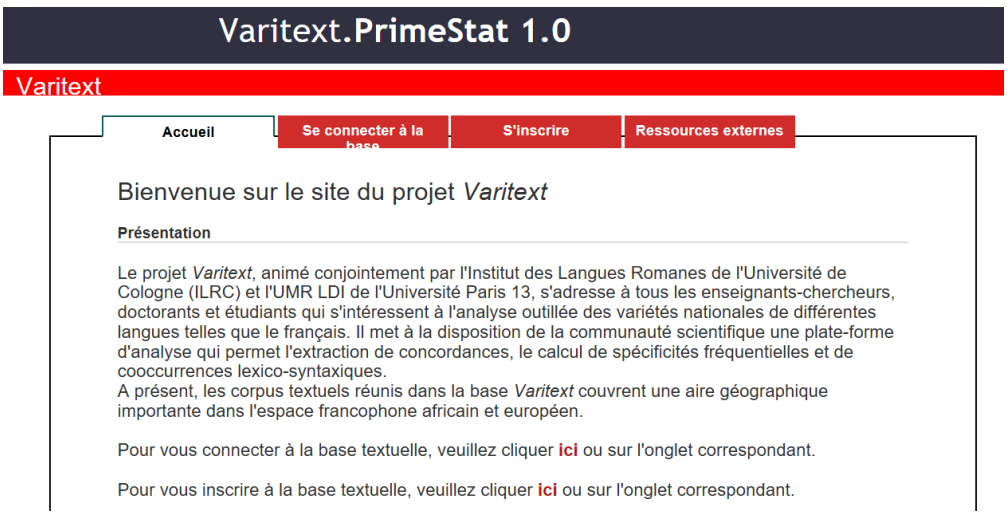

Abb. 1: Die webbasierte Plattform *Varitext*.

1 Das Projekt *Varitext* hat zum Ziel, Textressourcen, und zwar konkret linguistische Korpora, sowie eine online zugängliche Arbeitsumgebung für deren Analyse bereitzustellen. Dabei sei unter Textressourcen allgemein jede beliebige Art von digitaler Textsammlung verstanden, unter einem linguistischen Korpus hingegen speziell eine zu sprachwissenschaftlichen Zwecken zusammengestellte und aufbereitete digitale Textsammlung (s. u.). *Varitext* ist zugleich jene webbasierte Arbeitsumgebung (*platforme*), die Zugang zu den Korpora und den damit verknüpften Analysetools verschafft (siehe Abb. 1).

2 Ins Leben gerufen und getragen ist das Projekt laut Internetseite vom Romanischen Seminar der Universität zu Köln und vom *Laboratoire Lexique, Dictionnaire, Informatique* (LDI) der Universität Paris 13. Hauptsächlich gestaltet und betreut wird *Varitext* von Sascha Diwersy. Zielgruppe sind laut Webseite Forschende, Lehrende und Studierende, die sich für die Erforschung und Analyse sprachlicher Varietäten interessieren. Zum Entstehungskontext und zu den Bedingungen, unter denen das Projekt realisiert wird und wurde, werden leider keine Angaben gemacht. Ebenso finden sich bedauerlicherweise keine Angaben dazu, in welchem konkreten finanziellen und institutionellen Rahmen das Projekt weiter verfolgt wird. Beschreibungen, Hilfstexte und Arbeitsoberfläche sind bislang ausschließlich in französischer Sprache gehalten. *Varitext* ist einfach und frei online erreichbar und verwendbar; es genügen eine Registrierung und die Bestätigung durch das Team hinter *Varitext*.

3 Das Fernziel von *Varitext* ist es, Korpora zu (Standard-)Varietäten unterschiedlicher (und durchaus nicht nur romanischer) Sprachen zu sammeln und für die Analyse zugänglich zu machen: „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.“ (Diwersy 2014, 51). Bisher bietet die Plattform Zugang zu einem französischen Korpus, dem CoVaNa-FR: *Corpus des variétés nationales du français*. Die Integration eines Korpus spanischer Texte ist Diwersy (2014, 51) zufolge in Arbeit und die Zusammenstellung weiterer Korpora (Portugiesisch, Russisch, Arabisch) zumindest geplant. Zu Anpassungen der Plattform, z. B. sprachlicher oder technischer Art, die durch eine solche Erweiterung durch andere Sprachen notwendig werden können, sind derzeit noch keine Informationen verfügbar.

4 Über den genauen Inhalt des Korpus CoVaNa-FR, die Aufbereitung, die Annotation und die linguistische Einordnung des Korpus informiert ein Aufsatz Sascha Diwersys (2014) (leider nicht *Varitext* selbst, s. u.). CoVaNa-FR stellt ein schriftsprachliches Korpus dar und setzt sich (bislang) aus Texten der frankophonen Presse in Frankreich, Kanada, der Schweiz sowie mehreren afrikanischen Staaten zusammen, und zwar: Elfenbeinküste, Marokko, Algerien, der Demokratischen Republik Kongo, Mali, Kamerun, Senegal und Tunesien. Geplant ist eine Erweiterung um andere schriftsprachliche, nämlich fiktionale und akademische Texte. Mit einer solchen Ressource wird in der Tat, wie Sascha Diwersy (2014, 48) meint, ein wichtiges Desiderat innerhalb der Dokumentation und Erforschung des Französischen angegangen. Zwar ist das Französische durch diverse Korpora dokumentiert:[1] so bietet die Ressource *Frantext* z. B. einen „panchronen" (Pusch 2014, 183) Zugang zu französischen Texten mehrerer Jahrhunderte, bei denen es sich im Kern um literarische, essayistische, philosophische und wissenschaftliche Werke handelt; zum Teil lässt sich die Textsammlung als linguistisches Korpus im engeren Sinne nutzen. Eine Vielzahl größerer und kleinerer Projekte dokumentieren unterdessen das gesprochene Französisch, wie z. B. *Corpus de Référence du Français Parlé* oder *Corpus de Langue parlée en interaction* (CLAPI). Ein umfangreiches linguistisches Korpus zur Variation des (Standard-)Französischen im internationalen frankophonen Raum, das zudem einfach und kostenlos zugänglich ist, gibt es allerdings bislang, soweit ich sehe, nicht (siehe z. B. den Überblick von Pusch 2014).

5 Aus fast jedem der angegebenen frankophonen Staaten sind mindestens zwei verschiedene Angebote der nationalen Tagespresse vertreten (z. B. *Le Temps* und *La Tribune de Genève* für die Schweiz, *Fraternité Matin* und *Notre Voie* für die Elfenbeinküste). Erhoben wurden für jede Zeitung jeweils alle Ausgaben eines ganzen Jahres oder mehrerer Jahre. Schwerpunktmäßig sind die Jahre 2007 und 2008 vertreten, das heißt, dass fast jedes Presseteilkorpus[2] die Ausgaben dieser beiden oder eines dieser beiden Jahre umfasst. Insgesamt reicht die Zeitspanne, aus der Texte erhoben wurden, von 2003 bis 2009. Presse- und Ausgabenteilkorpora können also in dieser Hinsicht als vergleichbar angenommen werden. Derzeit umfasst CoVaNa-FR ungefähr 419.700.000 Token. Die Größe der Presseteilkorpora variiert zwischen 18,8 Mio. (Elfenbeinküste) und 53,5 Mio. Token (Kanada) (Diwersy 2014, 49), in Abhängigkeit von der Anzahl der verzeichneten Jahrgänge und sicher auch vom jeweiligen Angebot der Tageszeitungen.

6 Laut Diwersy (2014, 51) liegt das Korpus in XML vor und ist nach Subkorpus, Einzeltext, Absatz und Satz strukturiert. Da mit der *Open Corpus Workbench* gearbeitet wird, sind die Daten vertikalisiert, also in einer Art Tabelle mit XML-Tags, formatiert. Inwieweit die Annotationen kompatibel mit den Empfehlungen der *Text Encoding Initiative* sind, wird nicht expliziert. Das Korpus ist mit morphosyntaktischen (*Part-of-Speech*) und syntaktischen Informationen (*Parsing*) angereichert und darüber hinaus lemmatisiert. Die Aufbereitung der Texte ist mit Hilfe von *Connexor* Tools (Tagger) realisiert worden. *Varitext* liefert eine benutzungsfreundliche graphische Oberfläche und umfangreiche Auswertungsoptionen, für deren Aufbau die *Open Corpus Workbench, UCS toolkit* 0.6 sowie *R* verwendet wurden (Diwersy 2014, 49-52). Die Art der Aufbereitung und Strukturierung der Daten erlaubt eine differenzierte, individuelle Korpuskomposition und vielfältige Analysen und Auswertungen. Die Arbeitsumgebung ermöglicht Konkordanz- (KWIC-Konkordanzen), Frequenz- und Kookkurenzanalysen (s. u.).

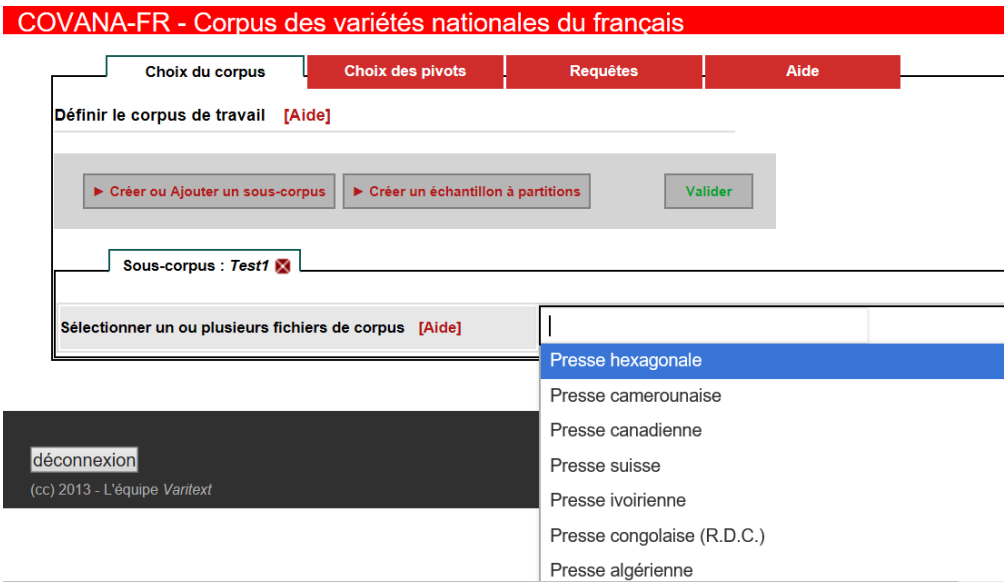

Abb. 2: Zusammenstellung des Untersuchungskorpus. Wahl der Presseteilkorpora.

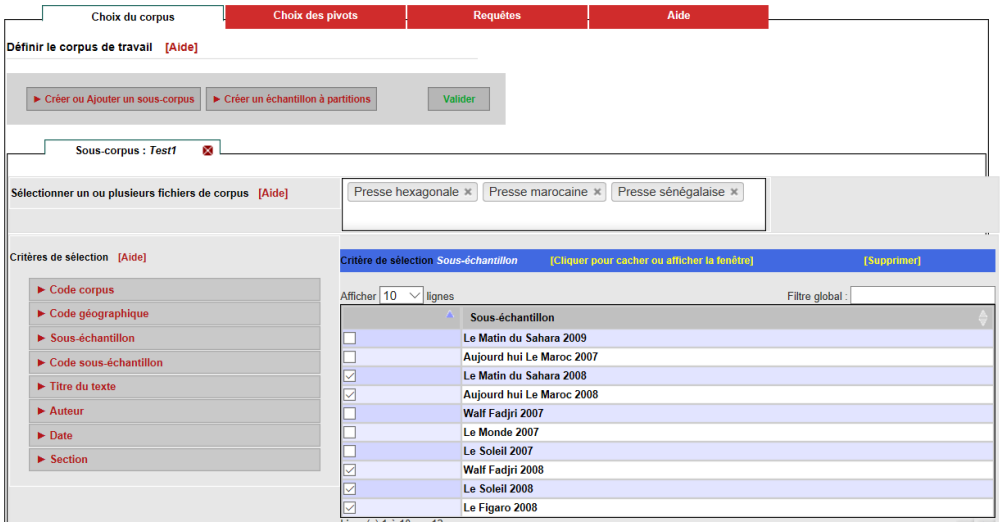

Abb. 3: Zusammenstellung des Untersuchungskorpus. Wahl der Ausgabenteilkorpora.

7 Um Analysen durchführen zu können, wird zunächst, unterstützt von der Arbeitsoberfläche, ein Untersuchungskorpus zusammengestellt (*créer ou ajouter un sous-corpus*, siehe Abb. 2 und 3).[3] Es kann jederzeit verfeinert und erweitert, aber leider nicht für einen späteren Gebrauch gespeichert werden. Über die Presse- und Ausgabenteilkorpora hinaus kann das Untersuchungskorpus nach einem genaueren Datum (*Date*), nach Themen bzw. Sparten (*section*) und AutorInnen (*Auteur*) präzisiert werden.[4] Höchst problematisch ist allerdings, dass der genaue Inhalt, die Größe und der zeitliche Kontext des Gesamtkorpus sowie seiner Subkorpora innerhalb der Arbeitsumgebung *Varitext* an keiner Stelle genau beschrieben werden. Die erste Zusammenstellung eines Untersuchungskorpus gleicht demnach einer Entdeckungsreise, weil erst im Verlaufe dieses Prozesses nach und nach erschlossen werden kann, wie CoVaNa-FR im Einzelnen aufgebaut ist.

Abb. 4: Eingabe und Differenzierung der gesuchten Wörter oder Ausdrücke.

Abb. 5: Präzisierung der Anfrage. Auswahl der Analysefunktion (KWIC-Konkordanz, Frequenz, Kookkurrenz), ggf. Regulierung der Ergebnisanzeige und des Kotextes.

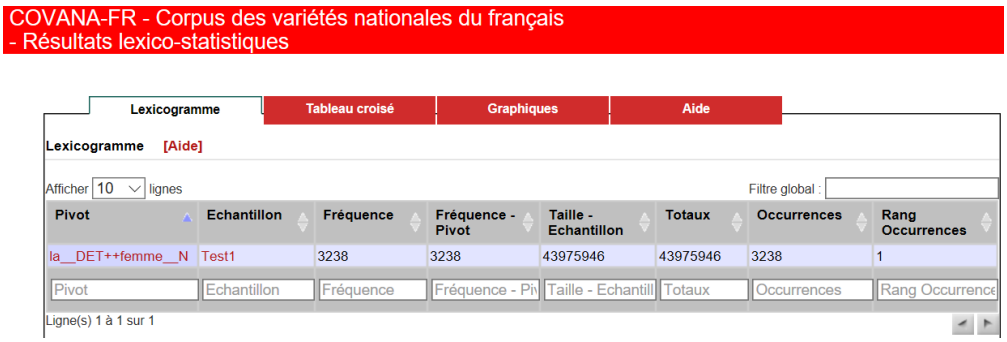

Abb. 6: Einfache Frequenzanalyse: Lexicogramme von *la+femme*.

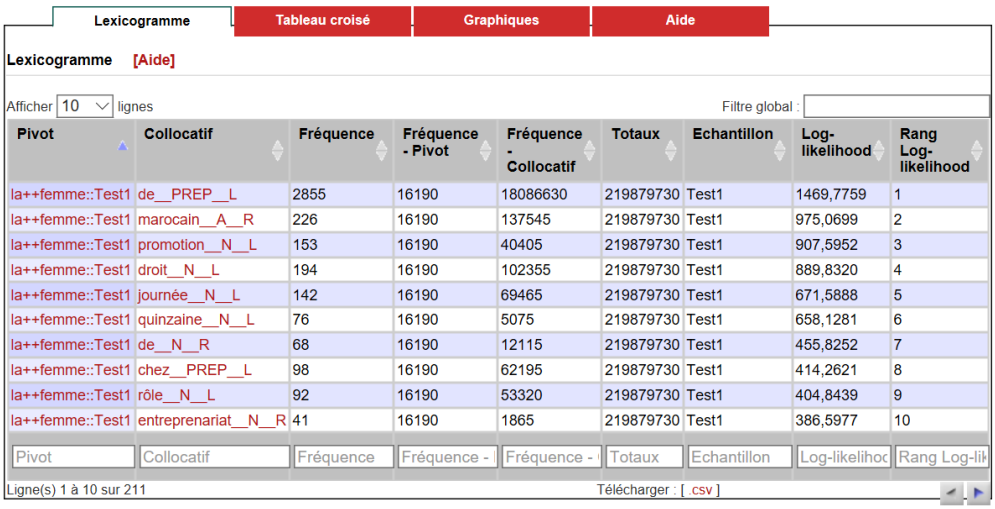

Abb. 7: Kookkurrenzanalyse zu „la femme“, zeigt häufigste linke (_L) und rechte (_R) Kollokatoren aus und gibt entsprechend den Einstellungen in der Anfrage statistische Werte aus.

8 Steht das (vorläufige) Untersuchungskorpus fest, können detaillierte Suchanfragen formuliert und entschieden werden, welche Art der Auswertung (KWIC-Konkordanzen, Frequenz-, Kookkurrenzanalysen) durchgeführt werden soll (siehe Abb. 4 und 5). Die Bearbeitung der zugrundeliegenden Textbasis (s. o.) erlaubt die Suche von exakten Wortformen, Lemmata, Kollokationen und die Verwendung von regulären Ausdrücken. Suchwörter und –phrasen können morphosyntaktisch spezifiziert werden, und zwar nach Wortart und nach syntaktischen Relationen und Funktionen (z. B. *sujet*, *attribut de l'objet*, *auxiliaire*). Eine Kombination verschiedener Kriterien ist möglich und ermöglicht komplexe Abfragen. Kookkurrenz- und Frequenzanalyse bieten umfangreiche statistische Auswertungen, die in verschiedenen Formaten (Lexikogramm, Tabelle, Graphik) dargestellt werden (siehe Abb. 6 und 7).

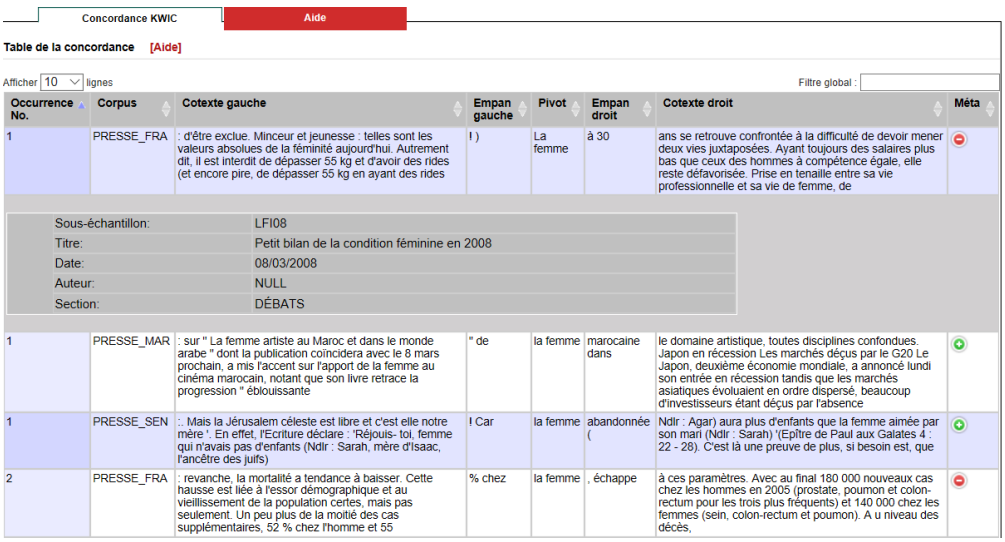

Abb. 8: Anzeige der Ergebnisse. Ergebnisse für KWIC-Konkordanzen zu *la+femme*.

9 Die Ausgabe der KWIC-Konkordanzen (möglich bis maximal 10.000 Ergebnisse) erfolgt in einer Tabelle, die das Suchwort, die unmittelbaren linken und rechten Nachbarn sowie den linksseitigen und rechtsseitigen Kotext (bis zu 50 Wörter) übersichtlich präsentiert. Zudem können ggf. die jeweiligen Presseteilkorpora, aus denen die Belege stammen, nachvollzogen und weitere Metadaten abgerufen werden (siehe Abb. 8). Die Inhalte der Tabellenspalten können zusätzlich sortiert und gefiltert werden. Die Ergebnisse sind in Form einer csv-Datei herunterladbar, stehen also bei Bedarf für eine weitere Verarbeitung zur Verfügung. Ein Zugriff auf die Volltexte ist jedoch nicht möglich. Die Zugriffsbeschränkungen verschiedener Art werden nur ungenau erläutert und begründet; so finden sich im Hilfetext lediglich recht allgemeine Hinweise dazu, dass die Ausgabe von Konkordanzen oder der Umfang des Kotextes

„pour des raison d'ordre juridique (et technique)" begrenzt ist. Deutlich gesagt wird, dass aus Gründen des Copyrights CoVaNa-FR nicht als komplette Ressource heruntergeladen werden kann, sondern ausschließlich in Teilen und über die graphische Benutzungsoberfläche zur Verfügung steht (Diwersy 2014, 49).

10 Es können hier nicht alle Funktionen ausführlich beschrieben werden: *Varitext* ist in jeder Hinsicht ein wertvolles Arbeitsinstrument für die Analyse der französischen Sprache. Die drei genannten Auswertungsfunktionen qualifizieren es insbesondere für Fragestellungen im Zusammenhang mit dem Lexikon, die Textzusammenstellung für den Vergleich innerhalb der Frankophonie. Das Hilfemenü (*Aide*) gibt zu allen Analysefunktionen eine verständliche und detaillierte Anleitung (in französischer Sprache), und die übersichtliche, teils anschauliche Darstellung der Ergebnisse erleichtert die Untersuchung und / oder Interpretation der Daten. So können die Möglichkeiten von *Varitext* auch von z. B. Studierenden mit vergleichsweise wenig Aufwand ausgeschöpft werden. Gleichzeitig bietet die Arbeitsplattform fortgeschrittenen Nutzenden ausreichend Differenzierungsmöglichkeiten bei der statistischen Auswertung.

11 Problematisch, da mehr als lückenhaft, ist leider insgesamt die Dokumentation und Beschreibung der Plattform *Varitext* und des Korpus CoVaNa-FR sowie seiner Voraussetzungen, konzeptionellen Verortung und genauen Zielstellung. Dies gilt in korpusmethodischer ebenso wie in variationslinguistischer Hinsicht. Zwar hat Sascha Diwersy in seinem 2014 veröffentlichten Beitrag zu den *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects* Vieles zu einer solchen Dokumentation beigetragen; jedoch ist fast nichts davon auf der Internetseite selbst nachvollziehbar und ist die Publikation dort auch nicht abgelegt oder zumindest verlinkt; es findet sich auch kein Hinweis darauf.

12 Aus (varietäten-)linguistischer Sicht fehlt eine Beschreibung und Einordnung von CoVaNa-FR im Rahmen der Plattform *Varitext* fast vollständig. Die Kurzvorstellung (*Accueil*, siehe Abb. 1) und die Nutzungsbedingungen (*Charte Varitext*) benennen als Inhalt der Textbasis geographische Varietäten der afrikanischen und europäischen Frankophonie: „la base *Varitext* couvrent une aire géographique importante dans l'espace francophone africain et européen"[5]. In diesem Zusammenhang findet der Begriff ‚variétés nationales' Verwendung, der auf der Webseite nicht weiter expliziert wird, sich aber auch im Namen der konkreten Textbasis wiederfindet: *COVANA-FR* –

Corpus des variétés nationales du français. Es ist davon auszugehen, und das bestätigt Diwersy (2014, 48-49), dass damit auf das Konzept der ‚nationalen Varietäten' sogenannter plurizentrischer Sprachen abgehoben wird. In dessen Sinne versteht z. B. Bernhard Pöll unter nationalen Varietäten „Ausprägungen der jeweiligen Standardvarietät in einem Gebiet, das entweder mit einem Staat oder mit einem staatsähnlichen Gebilde zusammenfällt" (Pöll 2000, 51). Dies vorausgesetzt, bezieht *CoVaNa-FR* sich also auf diejenigen Varietäten des Französischen, die in den jeweiligen frankophonen Staaten als „internes Standardfranzösisch" anerkannt und als solches zum hexagonalen Standard in Beziehung gesetzt werden können (Diwersy 2014, 48). Dies birgt jedoch verschiedene Probleme und sollte deshalb keinesfalls unkommentiert bleiben. Die fraglichen Punkte sollen hier angedeutet werden:

13 Erstens ist die Identifikation von Variations- und Nationengrenze, den der Begriff *variétés nationales* impliziert, nicht unproblematisch (siehe z. B. die Diskussion des Begriffs bei Reiffenstein 2001). Diwersy zufolge fokussiert die Korpuskonstruktion allerdings „elements of endonormative differentiation, i.e. the emergence of regionally specific norms compared to a supposed metropolitan standard variety of French" (Diwersy 2014, 48). Ob solche „regionalspezifischen Normen" mit dem Begriff ‚variété nationale' passend erfasst sind, wäre zu diskutieren. Diwersy findet hier eine Formulierung, die eigentlich für eine viel offenere begriffliche Konzeption steht und in der Lage wäre, Heterogenität innerhalb frankophoner Regionen zu integrieren. Die höchst komplexen und je eigenen Verhältnisse gerade in den multiethnischen und multilingualen afrikanischen Staaten lassen überhaupt die Einordnung des Französischen als ‚nationale' Sprache / Varietät in diesen ehemaligen Kolonien diskutabel erscheinen.

14 Zumal sich zweitens diese „nationalen Standardvarietäten" in ihren Funktionen für die Sprachgemeinschaften und in ihrer Reichweite durchaus stark unterscheiden können (cf. Bickel 2000); Standardsprachen bilden also nicht notwendigerweise ein einheitliches und auch kein selbsterklärendes Phänomen. In den frankophonen Staaten Afrikas hängt, wie Pöll zu entnehmen ist (1998, 95 ff.), die Reichweite des Französischen, zumal des Standardfranzösischen, nicht selten von Status, Bildung und ethnischer Zugehörigkeit der einzelnen Sprechenden ab. Bisweilen ist Französisch reine Schrift- und Verwaltungssprache. Das trifft so jedoch natürlich nicht für alle frankophonen Regionen zu. Die hiesige Konzeption von Standardsprache und -norm bezieht sich aber ganz offensichtlich allein auf die geschriebene Sprache und

identifiziert diese mit Standardvarietäten: „It should be obvious, then, that our present activities focus on diversifying the corpus resources, especially with regard to other written genres.“ (Diwersy 2014, 55). Dies wird in *Varitext* weder erwähnt noch begründet.

15 Drittens ergibt sich aus der (rudimentären) Beschreibung des Korpus auf der Internetseite einerseits und dem manifesten Inhalt des Korpus andererseits zunächst der Eindruck, dass hier stillschweigend vorausgesetzt wird, diese ‚nationalen (Standard-)Varietäten' würden durch die jeweilige überregionale Tagespresse repräsentiert. Das ist sicher keine abwegige Annahme, sie sollte aber dennoch begründet sein. In seinem Konferenzbeitrag verweist Diwersy (2014, 49) auf entsprechende Überlegungen zum Verhältnis zwischen Presse und Norm und betont auch, dass ein Ausbau des Korpus um weitere Textsorten vorgesehen ist:

> Due to its focus on endonormative differentiation, the CoVaNa-FR is less balanced with respect to genre than similar corpora for other languages […] 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 (…).
>
> (Diwersy 2014, 49.)

16 Es lässt sich diskutieren, ob Pressetexte Standardvarietäten „repräsentieren“ oder ob nicht vielmehr eine enge, wechselseitige Beziehung zwischen der Sprache überregionaler Presseerzeugnisse und sprachlichen Normen angenommen werden muss (siehe hierzu z. B. Burr 2004; Burkhardt et al. in Vorbereitung); und es muss gefragt werden, ob Pressetexte *allein* in der Lage sind für Standardvarietäten zu stehen, auch wenn das nur vorübergehend der Fall ist. Die Bezeichnung des bisherigen Korpus oder / und dessen Beschreibung sollten diesen Einschränkungen gegebenenfalls Rechnung tragen.[6]

17 Insgesamt wäre für die Nutzung der Ressource als linguistisches Korpus im engeren Sinne die Dokumentation

a) der linguistischen Verortung,
b) des Korpusdesigns in Bezug auf die linguistische Konzeption,

c) der Einschränkungen und (vorübergehenden) Lücken

unmittelbar im Rahmen der Präsentation des Korpus mehr als wünschenswert. Dies gilt ebenso für eine Beschreibung

d) der genauen Korpuszusammensetzung (Inhalt, Größe, Formate),
e) seiner Aufbereitung und Annotation und
f) der daraus resultierenden Potentiale.

Tatsächlich findet sich nur in den Nutzungsbedingungen von *Varitext* ein knapper Hinweis auf das XML-Format sowie auf PoS-Tagging, Lemmatisierung und Parsing. Dass das Korpus überhaupt und ausschließlich Pressetexte enthält, „entdeckt" die Nutzerin des Korpus erst, wenn sie beginnt, ein Untersuchungskorpus zusammenzustellen (s. o.).

18 Dass CoVaNa-FR nicht nur als linguistisches Korpus gedacht ist (dies vermittelt zumindest der Name), sondern als solches nutzbar ist, steht dabei außer Frage. Unter einem linguistischen Korpus im engeren Sinne verstehe ich in Anlehnung an Biber, Conrad und Reppen (1998, 246) sowie Burr (2004, 136-137) eine Sammlung elektronisch vorliegender Texte, die a) authentischen sozialen Kontexten entstammen, b) genau so, wie sie dort vorkamen, in das Korpus aufgenommen worden sind und c) mit Annotationen versehen sind, die mindestens die Zuordnung der Daten zu ihrem (zeitlichen, räumlichen usw.) Kontext erlauben. Ein linguistisches Korpus unterscheidet sich zudem von anderen ‚Textkollektionen' durch seine überlegte Zusammenstellung und Bearbeitung für linguistische Untersuchungen. In diesem Sinne definiert John Sinclair:

> A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.
>
> (Sinclair 2005, o. S.)

Das drückt sich im Falle der Textbasis von *Varitext* nicht nur in der Zusammenstellung nach externen Kriterien aus, wie es Sinclair fordert,[7] sondern eben auch in der umfangreichen linguistischen Annotation, die konkrete sprachwissenschaftliche Analysen stützt.

19 Ein wichtiges Kriterium ist darüber hinaus aber auch, dass das Korpus ausgewählte und beschreibbare Sprachausschnitte enthält, die geeignet sind, eine Sprache oder einen bestimmten Teil der Sprache (z. B. eine Varietät) in gewisser Hinsicht zu repräsentieren; die sich daraus ergebende Zusammensetzung bestimmt erheblich mit, welche Fragestellungen mit Hilfe des Korpus bearbeitet werden können: „The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research." (Biber, Conrad, and Reppen 1998, 246). Jedoch erfordert sowohl die Beurteilung der ‚Repräsentativität' eines Korpus in Bezug auf eine bestimmte Varietät als auch die Einschätzung, inwieweit die Ergebnisse von Datenanalysen verallgemeinerbar sein werden, als auch die Antizipation möglicher Fragestellungen, die mit Hilfe eines Korpus beantwortet werden könnten, gerade eine transparente, umfangreiche und methodisch wie konzeptionell fundierte Dokumentation.

20 Die Arbeitsplattform *Varitext* und das bisher enthaltene linguistische Korpus *Corpus des variétés nationales du français* stellen eine vielversprechende Ressource für die linguistische, und zwar insbesondere lexikalische und lexikologische Analyse französischer Varietäten in Europa, Afrika und Nordamerika dar. Zum jetzigen Zeitpunkt lässt sich das Korpus dabei in erster Linie als ‚Korpus frankophoner Pressesprachen' nutzen. Die Arbeitsumgebung *Varitext* bietet einen vergleichsweise benutzungsfreundlichen Zugang zum Textkorpus und stellt erhellende automatische Analysen sowie statistische Auswertungen bereit. Als dringend erforderlich erweist sich eine transparentere Darstellung und Beschreibung des Projekts, der Plattform und vor allem des Korpus; dies gilt sowohl in Hinsicht auf die methodische und inhaltliche Korpuskonstruktion als auch auf die sprachwissenschaftliche Verortung seines Inhalts. Eine solche Dokumentation stellt (einmal ganz abgesehen von ihrem Interessantheitswert) aus korpuslinguistischer Sicht die Voraussetzung für eine adäquate Nutzung der Ressource und die anschließende Bewertung der Ergebnisse dar.

Anmerkungen

1. Zum Französischen steht bislang kein ausgewogenes Referenzkorpus in dem Sinne zur Verfügung, wie es z.B. für das Deutsche mit dem DeReKo (Deutsches Referenzkorpus) des Instituts für deutsche Sprache der Fall ist. An einem *Corpus de*

référence du français contemporain arbeiten jedoch Dirk Siepmann, Christoph Bürgel und Sascha Diwersy (2016).

2. Das Gesamtkorpus CoVaNa-FR besteht aus Subkorpora auf der Ebene der Länder und ihrer jeweiligen Presse (*presse hexagonale*, *presse suisse*, *presse maroccaine*…). Diese wiederum bestehen aus den Subkorpora, die sich durch die Sammlung der Ausgaben eines bestimmten Jahres ergeben (z.B. *Le Monde* 2007). Ich bezeichne die Subkorpora der ersten Ebene als Presseteilkorpus. Die Subkorpora der zweiten Ebene als Ausgabenteilkorpus.

3. Alternativ kann für die Erstellung des Untersuchungskorpus die Variante *créer un échantillon à partition* gewählt werden. Die Varianten unterscheiden sich im Wesentlichen darin, dass die erste (créer un copus) einem top-down-Prozess entspricht: das Gesamtkorpus wird durch immer detailliertere Kriterien gegliedert. Unterdessen bietet die zweite einen schnelleren Zugang zu einem beliebig großen Teilkorpus, das nach einem präzisen Kriterium wie z. B. Datum gefiltert wurde.

4. Das ebenfalls mögliche Kriterium *Titre du Texte* liefert keine Ergebnisse. Weitere Filterkriterien sind *code géographique* und *code sous-échantillon*, da jedem Presseteilkorpus ein geographischer Code (z. B. CAN für Kanada) sowie ein Korpus-Code (z. B. PRESSE_CAN) zugeordnet ist. Jedes Ausgabenteilkorpus hat ebenfalls einen Code (z. B. DEV_CAN07 für *Le Devoir* 2007, Kanada).

5. Hier wurde Kanada vergessen.

6. Da sich das Korpus ausdrücklich auch an Studierende richtet, sollte das schon aus didaktischen Gründen geschehen.

7. Mit externen Kriterien ist die kommunikative Funktion von Texten gemeint, mit internen konkrete sprachliche Merkmale, die Sinclair (2005) zufolge jedoch nicht Auswahlkriterium, sondern vielmehr Gegenstand der Untersuchung sein sollen.

Bibliographie

Analyse et traitement informatique de la langue française. 2017. “Base textuelle FRANTEXT.“ Accessed 30.05.2017. http://www.frantext.fr/.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating language structure end use.* Cambridge: Cambridge University Press.

Bickel, Hans. 2000. “Deutsch in der Schweiz als nationale Varietät des Deutschen.“ *Sprachreport* (4): 21-27.

Burkhardt, Julia, Elena Potapenko, Rebekka Sierig, and Aramís Concepción Durán. In Vorbereitung. “Leipziger Frltz-Korpus: Das Korpus französischer und italienischer Zeitungssprache und seine Auszeichnung.“ *Romanische Studien.*

Burr, Elisabeth. 2004. “Das Korpus romanischer Zeitungssprachen in Forschung und Lehre.“ In *Romanistik und neue Medien: Romanistisches Kolloquium XVI*, edited by Wolfgang Dahmen, Günter Holtus, Johannes Kramer, Michael Metzeltin, Wolfgang Schweickard, and Otto Winkelmann, 133-162. Tübingen: Narr.

Laboratoire ICAR. 2014. “Corpus de Langue parlée en interaction.“ https://web.archive.org/save/_embed/http://clapi.ish-lyon.cnrs.fr/.

Diwersy, Sascha. 2014. “The Varitext platform and the Corpus des variétés nationales du français (CoVaNa-FR) as resources for the study of French from a pluricentric perspective.“ *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*: 48-57. doi: 10.3115/v1/W14-5306.

Pöll, Bernhard. 1998. *Französisch außerhalb Frankreichs: Geschichte, Status und Profil regionaler und nationaler Varietäten.* Tübingen: Niemeyer.

Pöll, Bernhard. 2000. “Plurizentrische Sprachen im Fremdsprachenunterricht (am Beispiel des Französischen)“ In *Normen im Fremdsprachenunterricht*, edited by Wolfgang Börner, Wolfgang and Klaus Vogel, 51-63. Tübingen: Narr.

Pusch, Claus D. 2014. “Les corpus romans contemporains.“ In *Manuels des languges romanes*, edited by Andre Klump, Johannes Kramer and Aline Willems, 173-196. Berlin / Boston: de Gruyter.

Reiffenstein, Ingo. 2010. “Das Problem der nationalen Varietäten. Rezensionsaufsatz zu Ulrich Ammon: Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten, Berlin/New York 1995.“ *Zeitschrift für Deutsche Philologie* 1: 78-89. Accessed 10.04.2017. https://www.zfdphdigital.de/ZFDPH.01.2001.078.

Siepmann, Dirk, Christoph Bürgel, and Sascha Diwersy. 2016. “Le Corpus de référence du français contemporain (CRFC), un corpus massif du français largement diversifié par genres.” In *SHS Web of Conferences* (27): 11002. doi: 10.1051/shsconf/20162711002.

Sinclair, John. 2005. “Corpus and Text: Basic Principles.“ In *Developing Linguistic Corpora: a Guide to Good Practice*, edited by Martin Wynne. Oxford: Oxbow Books https://web.archive.org/web/20170530110622/http://ota.ox.ac.uk/documents/creating/dlc/.

Factsheet

Resource reviewed	
Title	Varitext
Editors	Sascha Diwersy, Peter Blumenthal, Salah Mejri
URI	http://syrah.uni-koeln.de/varitext/
Publication Date	2013
Date of last access	15.08.2017

Reviewer	
Surname	Burkhardt
First Name	Julia
Organization	University of Leipzig
Place	Leipzig, Germany
Email	jburk (at) rz.uni-leipzig.de

General Information		
Bibliographic description	Can the text collection be identified in terms similar to traditional bibliographic descriptions (title, responsible editors, institution, date(s) of publication, identifier/address)? (cf. Catalogue 1.1)	yes
Contributors	Are the contributors (editors, institutions, associates) of the project documented? (cf. Catalogue 1.3)	yes
Contacts	Is contact information given? (cf. Catalogue 1.4)	yes
Aims		
Documentation	Is there a description of the aims and contents of the text collection? (cf. Catalogue 2.1)	no
Purpose	What is the purpose of the text collection? (cf. Catalogue 2.2)	Research, Teaching
Kind of research	What kind of research does the collection allow to conduct primarily? (cf. Catalogue 3.1.8)	Quantitative research

Self-classification	How does the text collection classify itself (e.g. in its title or documentation)? (cf. Catalogue 2.3)	Corpus, Database
Field of research	To which field(s) of research does the text collection contribute? (cf. Catalogue 2.2)	Linguistics
Content		
Era	What era(s) do the texts belong to? (cf. Catalogue 2.5)	Contemporary
Language	What languages are the texts in? (cf. Catalogue 2.5)	French
Types of text	What kind of texts are in the collection? (cf. Catalogue 2.5)	Newspaper/journal articles
Additional information	What kind of information is published in addition to the texts? (cf. Catalogue 2.5)	none
Composition		
Documentation	Are the principles and decisions regarding the design of the text collection, its composition and the selection of texts documented? (cf. Catalogue 3.1.1-3.1.3)	no
Selection	What selection criteria have been chosen for the text collection? (cf. Catalogue 3.1)	Language, Country, Genre
Size		
Texts/records	How large is the text collection in number of texts/ records? (cf. Catalogue 3.1.4)	> 1000
Tokens	How large is the text collection in number of tokens? (cf. Catalogue 3.1.4)	> 10 Mio.
Structure	Does the text collection have identifiable sub-collections or components? (cf. Catalogue 3.1.5)	yes
Data acquisition and integration		
Text recording	Does the text collection record or transcribe the textual data for the first time? (cf. Catalogue 3.1.6)	no
Text integration	What kind of material has been taken over from other sources? (cf. Catalogue 3.1.6)	Full texts, Metadata

Quality assurance	Has the quality of the data (transcriptions, metadata, annotations, etc.) been checked? (cf. Catalogue 3.1.7)	unknown
Typology	Considering aims and methods of the text collection, how would you classify it further? For definitions please consider the help-texts. (cf. Catalogue 3.1.8)	Corpus
Data Modelling		
Text treatment	How are the textual sources represented in the digital collection? (cf. Catalogue 3.2.1)	other:
Basic format	In which basic format are the texts encoded? (cf. Catalogue 3.2.4)	XML
Annotations		
Annotation type	With what information are the texts further enriched? (cf. Catalogue 3.2.2)	Linguistic annotations, Structural information
Annotation integration	How are the annotations linked to the texts themselves? (cf. Catalogue 3.2.2)	Stand-off
Metadata		
Metadata type	What kind of metadata are included in the text collection? (cf. Catalogue 3.2.3)	Descriptive, Structural
Metadata level	On which level are the metadata included? (cf. Catalogue 3.2.2)	Individual texts
Data schemas and standards		
Schemas	What kind of data/metadata/annotation schemas are used for the text collection? (cf. Catalogue 3.2.4)	Project specific schema
Standards	Which standards for text encoding, metadata and annotation are used in the text collection? (cf. Catalogue 3.2.4)	TEI
Provision		
Accessability of the basic data	Is the textual data accessible in a source format (e.g. XML, TXT)? (cf. Catalogue 4.1)	no
Download	Can the entire raw data of the project be downloaded (as a whole)? (cf. Catalogue 4.2)	no

Technical interfaces	Are there technical interfaces which allow the reuse of the data of the text collection in other contexts? (cf. Catalogue 4.2)	none
Analytical data	Besides the textual data, does the project provide analytical data (e.g. statistics) to download or harvest? (cf. Catalogue 4.3)	yes
Reuse	Can you use the data with other tools useful for this kind of content? (cf. Catalogue 4.4)	no
User Interface		
Interface provision	Does the text collection have a dedicated user interface designed for the collection at hand in which the texts of the collection are represented and/or in which the data is analyzable? (cf. Catalogue 5.1)	yes
User Interface questions		
Usability	From your point of view, is the interface of the text collection clearly arranged and easy to navigate so that the user can quickly identify the purpose, the content and the main access methods of the resource? (cf. Catalogue 5.3)	no
Acces modes		
Browsing	Does the project offer the possibility to browse the contents by simple browsing options or advanced structured access via indices (e.g. by author, year, genre)? (cf. Catalogue 5.4)	yes
Fulltext search	Does the project offer a fulltext search? (cf. Catalogue 5.4)	yes
Advanced search	Does the project offer an advanced search? (cf. Catalogue 5.4)	yes
Analysis		
Tools	Does the text collection integrate tools for analyses of the data? (cf. Catalogue 5.5)	yes
Customization	Can the user alter the interface in order to affect the outcomes of representation and analysis of the text collection (besides basic search functionalities), e.g. by applying his or her own queries or by choosing analysis parameters? (cf. Catalogue 5.5)	yes

Visualization	Does the text collection provide particular visualizations of the data? (cf. Catalogue 5.6)	Charts
Personalization	Is there a personalisation mode that enables the users e.g. to create their own sub-collections of the existing text collection? (cf. Catalogue 5.7)	no
Preservation		
Documentation	Does the text collection provide sufficient documentation about the project in general as well as about the aims, contents and methods of the text collection? (cf. Catalogue 6.1)	no
Open Access	Is the text collection Open Access? (cf. Catalogue 6.2)	yes
Rights		
Declared	Are the rights to (re)use the content declared? (cf. Catalogue 6.2)	yes
License	Under what license are the contents released? (cf. Catalogue 6.2)	No explicit license / all rights reserved
Persistent identification and addressing	Are there persistent identifiers and an addressing system for the text collection and/or parts/objects of it and which mechanism is used to that end? (cf. Catalogue 6.3)	Persistent URLs
Citation	Does the text collection supply citation guidelines? (cf. Catalogue 6.3)	yes
Archiving of the data	Does the documentation include information about the long term sustainability of the basic data (archiving of the data)? (cf. Catalogue 6.4)	no
Institutional curation	Does the project provide information about institutional support for the curation and sustainability of the project? (cf. Catalogue 6.4)	no
Completion	Is the text collection completed? (cf. Catalogue 6.4)	yes
Personnel		
Editors	Peter Blumenthal Salah Mejri	
Designers	Sascha Diversy	