

TEXTDATEN



ROMANISTIK

# Textdaten Romanistik

Ein Wiki zur Arbeit mit digitalen Ressourcen in den  
romanischen Sprachwissenschaften

# Blended Learning-Konzept

## Präsenzveranstaltungen

Innerhalb und außerhalb von Seminaren

Zielgruppenspezifisch: Teilbereich oder allgemeine Einführung

Fortsetzung im Selbststudium

## Sprechstundenunterstützung

Unterstützung der Sprechstunde: Hausarbeiten, Abschlussarbeiten

Fortsetzung im Selbststudium

Rückfragen in der Sprechstunde

## Nutzer von außen

Selbststudium

# TEXT



# DATEN

Seiten / Textdaten Romanistik Home

Selbstlernmaterialien zum Wiki

Bearbeiten Beobachten Teilen Extras ▾

Der bewusst weit gefasste Begriff *Textdaten* verweist auf die diversen Arten von *Texten* - Korpora, historische und aktuelle Literatur, Wörterbücher, Sprachatlanten, Sekundärliteratur - die als Informationsquelle oder als direkter Untersuchungsgegenstand zur Klärung einer Fragestellung beitragen können.

Von *Daten* ist die Rede, da es sich um digitalisierte Texte handelt, die dem Benutzer online zugänglich sind.

Textdaten können mithilfe entsprechender Software digital analysiert und weiterverarbeitet werden.

Damit bilden sie eine Datengrundlage, auf der sich sprachliche Merkmale differenziert und statistisch aussagekräftig untersuchen lassen.

Aufschluss über weiter zurückliegende Sprachstadien geben und damit auch zum Objekt linguistischer Untersuchungen werden können. Sie stehen in Online-Bibliotheken und Datenbanken zur Verfügung.

In den Bereichen [Wörterbücher](#) und [Sprachatlanten](#) kann der Nutzer sich über lexikalische Besonderheiten einzelner Sprachen, Sprachstufen und Varietäten informieren und sich ein Bild von der regionalen Ausprägung der Sprachen der Romania machen.

Zur Unterstützung der selbständigen Informationssammlung im Internet und der Erschließung unterstützender Forschungsliteratur verweist das Wiki Textdaten Romanistik im entsprechenden Bereich auf einschlägige bibliographische, biographische und enzyklopädische [Rechercheportale](#).

Wer sich an die aktive Arbeit mit den Textdaten wagen will, und vielleicht sogar selbst Korpora erstellen oder Texte zur Analyse aufbereiten möchte, findet im Bereich [Tools](#) diverse gängige Programme zur Transkription und Annotation.

# Welche Ziele werden verfolgt?

- Bereitstellen von Sprachdaten und Analysetools
- Heranführen an die Arbeit von Textdaten mithilfe praxisorientierter Manuals
- Liste von linguistisch relevanten Konventionen
- Förderung der computergestützten Arbeit mit Daten
- Entlastung von Seminaren und Sprechstunden: Antworten zu häufigen Fragen
- Einbindung der Studierenden: Evaluation und Übungen

# Inhalte des Wikis

- Das Wiki selbst enthält keine Textdaten, Ressourcen werden
  - verlinkt
  - geordnet
  - beschrieben
- Das Wiki enthält weiterhin
  - Links zu Programmen, die der Arbeit mit den Textdaten dienen
  - Anleitungen für diese Programme
  - Selbstlernmaterialien
  - Übungen

# Wie findet man im Wiki, was man sucht?

Ein detailliertes **Schlagwortsystem** ermöglicht die thematische Suche nach benötigten Ressourcen:

- ▶ Art der Ressource (*corpus, database, dictionary, tool...*)
- ▶ Sprache (*portuguese, catalan, latin, anglo-norman...*)
- ▶ Textsorte (*interviews, journalism, literature, theatre, language\_and\_law...*)
- ▶ Historische Einordnung (*1300, old\_french, renaissance, middle\_ages, diachrony...*)
- ▶ Sprachliche Besonderheiten (*dialects, child\_language, written\_language...*)
- ▶ Aufbereitung/Analysemöglichkeiten (*part\_of\_speech, lemmatized...*)

# Häufig gesuchte Labels

## Labels

1000 1100 1200 1300 1400 1500 1600 1700 1800 **1900 2000** 800 900 africa anglo-norman atlas audio bibliography biographical\_information brazil  
canada catalan child\_language **contemporary\_resources corpus database diachrony** dialects dictionary **digital\_edition**  
early\_modern\_spanish early\_printed\_book encyclopedia english europe european\_union francophonie francoprovençal français\_langue\_étrangère **french**  
frequency frequency\_list german germanic\_languages grammar **historical\_resources** history interviews **italian key\_word\_in\_context**  
language\_acquisition language\_and\_law language\_comparison language\_for\_special\_purposes language\_in\_politics latin latin\_america **lemmatized** link\_list  
**literature** manual manuscript media middle\_ages middle\_french multilingual non\_european\_languages old\_age old\_french old\_spanish parsing  
**part\_of\_speech** phonetics phonology portuguese press prosody regional\_languages renaissance research\_portal romance\_languages romanian  
scientific\_language second\_language\_acquisition sentiment\_analysis slavic\_languages sociolinguistics spanish **spoken\_language** syntactic\_annotation  
theatre tokenized tool **transcription** translation varieties video web\_as\_corpus **written\_language** youth\_language

# Bereiche: 1. Linguistische Korpusarbeit

- Mithilfe von Korpora kann man sehr **große Datenmengen** zuverlässig analysieren
- Sprachliche Merkmale werden computerlesbar und **zählbar** gemacht
- **Tools und Anwendungen** helfen bei der Annotation und Auswertung von Texten.
- Auf Basis der **Häufigkeit charakteristischer Merkmale** ist eine Analyse von grammatischen Strukturen, die Beschreibung von Sprachregistern, Varietäten, Sprachwandel usw. möglich



# Bereiche: 1. Linguistische Korpora

- ▼ Französische Korpora
  - [Actes royaux de Poitou \(1302-1464\)](#)
  - [Anglo-Norman Source Texts](#)
  - **Base de Français Médiéval**
  - [Bibliothèque bleue de Troyes](#)
  - [Corpus d'articles de linguistique issus de la revue "Sciences Humaines"](#)
  - [Corpus de Français Parlé Parisien des années 2000 \(CFPP2000\)](#)
  - [Corpus de la littérature médiévale: des origines au XVe siècle](#)
  - [Corpus de la littérature narrative du moyen âge au XXe siècle](#)
  - [Corpus de Langues Parlées en Interaction \(CLAPI\)](#)

Seiten / ... / Französische Korpora

Extras ▾

## Base de Français Médiéval

Umfangreiches Korpus zum Alt- und Mittelfranzösischen. Breites Spektrum an Genres und Dialekten. Morphosyntaktisch annotiert.

<b>Sprache</b>	Französisch
<b>Sprachstufe</b>	Alt- und Mittelfranzösisch
<b>Sprachliche Realisierung</b>	schriftlich
<b>Umfang</b>	126 Volltexte, ca. 3,5 Mio. Wörter
<b>Medium</b>	breites Spektrum digitalisierter mittelalterlicher Texte, sortiert nach Genres (Literatur, Geschichte, Geographie, Rechtstexte, Wissenschaft...) und Dialekten
<b>Geographischer Ursprung</b>	Frankreich
<b>Zeitliche Einordnung</b>	9.-15. Jh.
<b>Form der Daten</b>	Gesamttexte, online durchsuchbar und z.T. in PDF-Format herunterladbar; detaillierte Metadaten verfügbar
<b>Format</b>	XML, PDF
<b>Annotation</b>	Lemmatisiert, morphosyntaktisch annotiert
<b>Mögliche Suchabfragen</b>	Suche nach Lemmata und morphosyntaktischen Kategorien nach Download der Software TXM
<b>Quelle/Herausgeber</b>	ENS/CAR/Universität de Lyon
<b>Nutzungsvoraussetzungen</b>	freier Zugang zum Lesen und Download der Volltexte; Registrierung für weitere Funktionen nötig
<b>Link</b>	<a href="http://bfm.ens-lyon.fr/">http://bfm.ens-lyon.fr/</a>

Schlagwörter (Labels)



Infos und Link zum Korpus

[corpus](#) [french](#) [old\\_french](#) [written\\_language](#) [diachrony](#)  
[middle\\_ages](#) [historical\\_resources](#)

# Bereiche: 2. Rechercheportale



Textdaten Romanistik

SEITENHIERARCHIE

- > Selbstlernmaterialien zum Wiki
- > Korpora und Textdatenbanken
- > Digitale Editionen
- Rechercheportale**
  - ▼ Allgemeine Portale
    - **MLA International Bibliography**
    - Open Access in den Philologien
    - **Romanische Bibliographie Online**

Seiten / ... / Allgemeine Portale

## Romanische Bibliographie Online

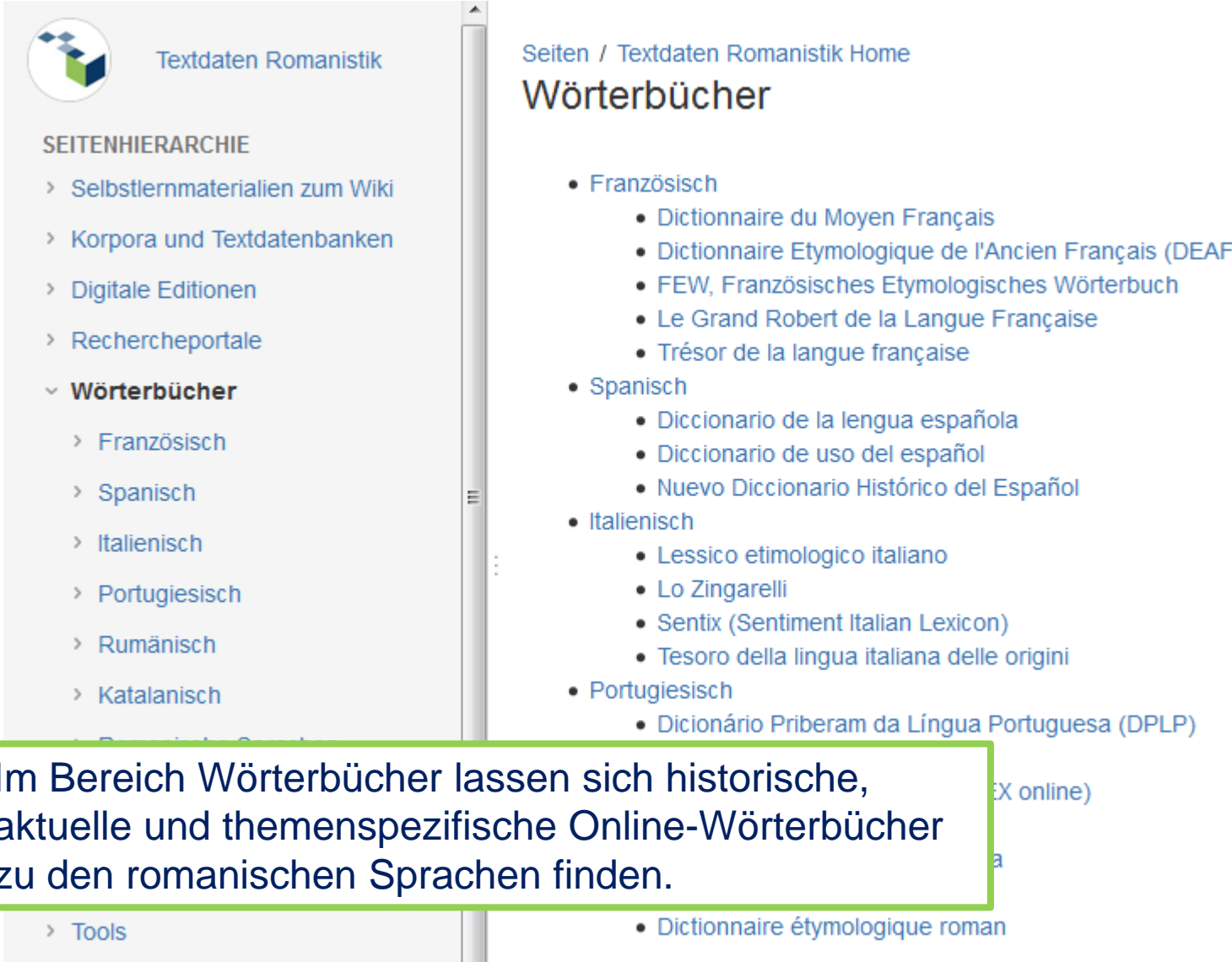
Umfassende romanistische Fachbibliographie des De Gruyter-Verlages. 450 000 Einträge.

Zugang über das Datenbanksystem der FU: [http://www.ub.fu-berlin.de/digibib\\_neu/datenbank/metalib/titel/KOB19716.html](http://www.ub.fu-berlin.de/digibib_neu/datenbank/metalib/titel/KOB19716.html)

romance\_languages romanian french  
italian spanish portuguese  
bibliography research\_portal

Unter den Begriff Rechercheportale fallen Online-Bibliographien, Suchmaschinen und Enzyklopädien, die bei der Suche nach Fachliteratur und Fachwissen helfen.

# Bereiche: 3. Wörterbücher



The screenshot shows the 'Textdaten Romanistik' website. The sidebar on the left contains a navigation menu under 'SEITENHIERARCHIE' with the following items: 'Selbstlernmaterialien zum Wiki', 'Korpora und Textdatenbanken', 'Digitale Editionen', 'Rechercheportale', 'Wörterbücher' (expanded), 'Französisch', 'Spanisch', 'Italienisch', 'Portugiesisch', 'Rumänisch', 'Katalanisch', and 'Tools'. The main content area is titled 'Seiten / Textdaten Romanistik Home' and 'Wörterbücher'. It lists dictionaries for four languages: French, Spanish, Italian, and Portuguese. A green box highlights a text block at the bottom of the page.

Textdaten Romanistik

Seiten / Textdaten Romanistik Home

## Wörterbücher

- Französisch
  - Dictionnaire du Moyen Français
  - Dictionnaire Etymologique de l'Ancien Français (DEAF)
  - FEW, Französisches Etymologisches Wörterbuch
  - Le Grand Robert de la Langue Française
  - Trésor de la langue française
- Spanisch
  - Diccionario de la lengua española
  - Diccionario de uso del español
  - Nuevo Diccionario Histórico del Español
- Italienisch
  - Lessico etimologico italiano
  - Lo Zingarelli
  - Sentix (Sentiment Italian Lexicon)
  - Tesoro della lingua italiana delle origini
- Portugiesisch
  - Dicionário Priberam da Língua Portuguesa (DPLP)

Im Bereich Wörterbücher lassen sich historische, aktuelle und themenspezifische Online-Wörterbücher zu den romanischen Sprachen finden.

• Dictionnaire étymologique roman

# Bereiche: 4. Sprachatlanten

- › Rechercheportale
- › Wörterbücher
- ▼ Sprachatlanten
  - Frankreich
  - Spanien
  - ▼ **Italien**
    - AIS - Sprach- und Sachatlas Italiens und der Südschweiz
    - ALI - Atlante linguistico italiano
    - ALiQuot - Atlante della lingua italiana quotidiana
    - AsiCa - Atlante sintattico della Calabria
    - VIVALDI - Vivaio Acustico delle Lingue e dei Dialetti d'Italia

Seiten / Textdaten Romanistik Home / Sprachatlanten

## Italien

[AIS - Sprach- und Sachatlas Italiens und der Südschweiz](#)

[ALI - Atlante linguistico italiano](#)

[ALiQuot - Atlante della lingua italiana quotidiana](#)

[AsiCa - Atlante sintattico della Calabria](#)

[VIVALDI - Vivaio Acustico delle Lingue e dei Dialetti d'Italia](#)

Online-Sprachatlanten informieren mithilfe von Karten und Audiomaterial über sprachliche und dialektale Besonderheiten bestimmter Regionen.

Compendium - ein Enterprise Wiki, bereitgestellt von



Wikis-Home

Nutzungsbeding

# Bereiche: 5. Tools

- › Rechercheportale
- › Wörterbücher
- › Sprachatlanten
- › Manuals
- ▼ Tools
  - DARIAH GeoBrowser
  - › Konkordanzprogramme
    - Morphalou
    - MORPH-IT!
    - Praat
    - **Statistikprogramm R**
    - Transcriber
    - TreeTagger
    - Stuttgart Corpus Workbench

Seiten

/ [Textdaten Romanistik Home](#)

/ [Tools](#)

## Statistikprogramm R

Statistikprogramm, mit dem aber auch Textdaten eingelesen und analysiert werden - daher gut geeignet für die Korpusanalyse.

Frei verfügbar unter

<http://www.r-project.org/>

Linguistische Analyse in R: Webseite von [Harald Baayen](#) und Einführungen von [Stefan Gries](#)

Erste Schritte in R (PDF zum Download): [Einführung von Heike Zinsmeister](#)

First steps in R (PDF zum Download): [Einführung von Joost van de Weijer Dylan Glynn](#)

👍 [Gefällt mir](#) Sei der Erste, dem dies gefällt.

[tool](#) [statistics](#) [corpus](#) [frequency](#) ✎

Im Bereich Tools befindet sich eine Liste von Programmen, die für die Arbeit mit Textdaten gebraucht werden. Zu jedem Programm gibt es eine Kurzbeschreibung und den Link zur Installation.

# Bereiche: 6. Manuals

## SEITENHIERARCHIE

- › Korpora und Textdatenbanken
- › Digitale Editionen
- › Rechercheportale
- › Wörterbücher
- › Sprachatlanten
- ✓ Manuals
  - › Audiotranskription
    - Praat Kurzanleitung
    - Tastaturfreundliche phonetische Zeichen: SAMPA statt IPA
    - **Transcriber Kurzanleitung**
  - › Datenkodierung und XML
  - Digitale Textedition mit TEI
  - Programmieren in der Linguistik
  - › Selbstlernmaterialien zum Wiki
  - › Wissenschaftliches Arbeiten, allgemein
- › Tools

## Arbeiten mit Transcriber

Das Programm ist im Internet frei verfügbar (<http://sourceforge.net/projects/trans/files/transcriber/1.5.1/>). Installation entsprechend den Anweisungen in der *Readme*-Datei.

Das Programm erzeugt ein Icon auf dem Desktop. Wenn es angeklickt wird, startet das Programm.

Zunächst kann eine eigene Tondatei (Format \*.wav) geladen werden: "File", "New trans", im Explorerfenster Ablageort der entsprechenden \*.wav-D:

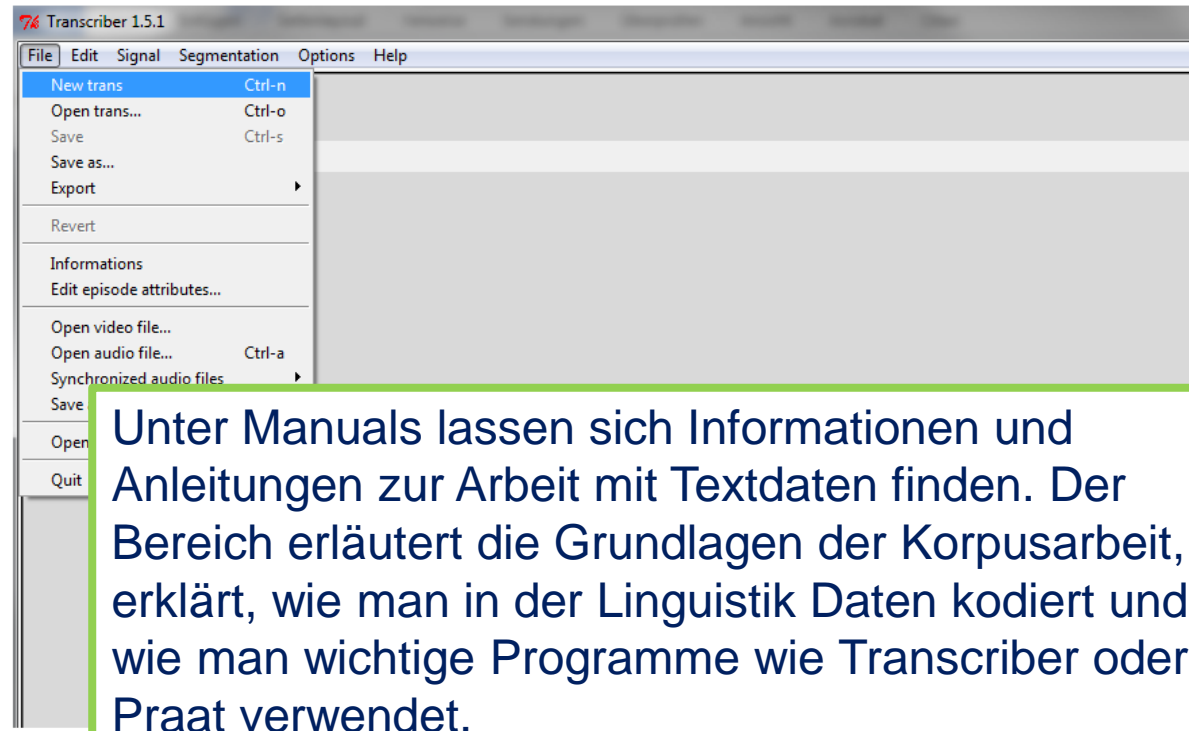


Fig. 1 Erstellen einer neuen Transkription

# Bereiche: 7. Wissenschaftliche Methoden



Textdaten Romanistik

## SEITENHIERARCHIE

- › Selbstlernmaterialien zum Wiki
- › Korpora und Textdatenbanken
- › Digitale Editionen
- › Rechercheportale
- › Wörterbücher
- › Sprachatlanten
- › Manuals
- › Tools
- › **Wissenschaftliche Methoden**
- Über Textdaten Romanistik



## ▼ **Wissenschaftliche Methoden**

- Arbeitstechniken für Romanisten
- Experimentelle Methoden
- Leipzig Glossing Rules
- Wissenschaftliche Standards (DFG)

# Bereiche: 7. Wissenschaftliche Methoden

- › Wörterbücher
- › Sprachatlanten
- › Manuals
- › Tools
- ▼ Wissenschaftliche Methoden
  - Arbeitstechniken für Romanisten
  - Experimentelle Methoden
  - Leipzig Glossing Rules
  - **Wissenschaftliche Standards (DFG)**

[Seiten](#) / [Textdaten Romanistik Home](#) / [Wissenschaftliche Methoden](#)

## Wissenschaftliche Standards (DFG)

**Allgemein: Verhindern von Plagiaten, Standards des Zitierens, ...**

Empfehlungen zum Einhalten der Regeln "Guter wissenschaftlicher Praxis"

**Empfehlungen der DFG zum Aufbau von Korpora**

Empfehlungen der DFG zu Datenformaten

Empfehlungen der DFG zu rechtlichen Aspekten

In diesem Bereich befinden sich Informationen zu den Grundlagen wissenschaftlichen Arbeitens: dazu, wie man erfolgreich eine wissenschaftliche Arbeit schreibt und welche Regeln beim Durchführen von Experimenten, beim Zitieren oder Glossieren gelten.



# Relevante Literatur

- Gerstenberg, Annette (<sup>2</sup>2013): *Arbeitstechniken für Romanisten. Eine Anleitung für den Bereich Linguistik*. Berlin/Boston: De Gruyter (speziell Kap. 6).  
HSK 29.1 = Lüdeling, Anke / Kytö, Merja (eds.) (2008): *Corpus Linguistics. An International Handbook*. Berlin/New York: De Gruyter.
- Lemnitzer, Lothar / Zinsmeister, Heike (eds.) (2006): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.
- Pusch, Claus D. / Raible, Wolfgang (2002): *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache*. Tübingen: Narr (ScriptOralia, 126).
- Barbera, Manuel / Corino, Elisa / Onesti, Cristina (eds.) (2007): *Corpora e linguistica in rete*. Perugia: Guerra.
- Bilger, Mireille (1996): „Corpus de Portugais et d'Espagnol“. In: *Recherches sur le français parlé* 1/2, 124-130.
- Freddi, Maria (2014): *Linguistica dei corpora*. Roma: Carocci (Bussole, 490).
- Williams, Geoffrey (ed.) (2005): *La linguistique de corpus*. Rennes: Presses universitaires de Rennes.