第5回　文化審議会国語分科会言語資源小委員会
2024年12月26日（木）

言語資源小委(R6. 12. 26)
資 料 2

ドイツ語コーパスの整備状況とその活用例

今道　晴彦
（広島大学人間社会科学研究科）

目次

1 ドイツ語コーパスの整備状況
- ドイツ語のコーパス開発
- ドイツ語圏の主要コーパス

2 ドイツ語コーパスの活用例
- ドイツの場合
- オーストリアの場合

1 ドイツ語コーパスの整備状況

ドイツ語のコーパス開発

- ドイツ・マンハイムのドイツ語研究所（IDS）で1964年の設立年にPaul GrebeとUlrich Engelにより**Mannheimer Korpus**プロジェクトが開始される
- 背景（Geyken, 2007；Teubert & Belica, 2014）
  - →過去の清算と啓蒙主義的理想への回帰
  - →言語研究におけるデータ収集の必要性
  - →電子計算機の普及

※ただし，世界初の「ブラウンコーパス」が発表された事実を当時IDSの理事会がどれほど認識していたかは不明とされる

Mannheimer Korpus（MK1）

- 1967年に完成
- 総語数：約220万語
- テキスト数：293
- 年代：1950年～1967年
- ジャンル：小説（Heinrich Böll，Günter Grass，Thomas Mann などの著名人の代表作が含まれる），回想録，学術テキスト，新聞・雑誌
- データをパンチカードに記録（IDS）→機械可読化（ダルムシュタット工科大学計算機センター）→修正作業（IDS）
- ただし，完成後に代表性・均衡性の問題が指摘される

Mannheimer Korpus 2（MK2）

- MK1の限界を補うため**Mannheimer Korpus 2**が構想される
- 総語数：約30万語
- テキスト数：52
- 年代：1949年，1952年，1960年～1974年
- ジャンル：政令，法令，決議，使用説明書，娯楽小説，教科書，ニュース
- MK1では無視されたテキストも収録される

LIMAS-Korpus

- **LIMAS**=**Li**nguistik und **Ma**schinelle **S**prachbearbeitung（言語学と機械言語処理）
- 1973年に完成
- 総語数：約100万語
- テキスト数：500（1テキストあたり2000語）
- 年代：1970年～1971年
- ジャンル：33種
- ブラウンコーパスをモデルとする
- ドイツ国立図書館で作成された書誌情報（Deutsche Nationalbibliographie）に基づき，テキストが選定される

ドイツ語のコーパス開発（続き）

- ただし，コーパスの利用は当時限定的であった
- 例えば動詞が属格目的語と前置詞目的語の両方を取る事例を調査する場合，煩雑な作業が求められた
  1）可能な動詞とすべての語形をリストアップし，
  2）プログラムを書いて，パンチカードに入力し
  3）近隣のダルムシュタット工科大学の計算機で解析し，
  4）専門家による出力されたコンコーダンスの修正を経て，
  5）研究者が手動で属格目的語と前置詞目的語を分類する必要があった

ドイツ語のコーパス開発（続き）

- 1969年にIDSで計算機システムが導入される
- しかし，1980年代に至るまで十分に活用されなかった
- Teubert & Belica（2014:302）

Das Korpus wurde oft eher als Steinbruch für Belegstellen und nicht im Sinne einer systematischen Datenanalyse genutzt.（コーパスは体系的なデータ分析というより事例を取り出すための採石場として利用されることが多かった。）

ドイツ語のコーパス開発（続き）

- その後，IDSで構築された種々のコーパスが新しいプラットフォームに統合され，書き言葉のドイツ語のアーカイブを構築するために，継続的なテキスト収集がなされる
- このアーカイブは後述のドイツ語書き言葉参照コーパスの**DeReKo**として継承される
- また，アーカイブ化に合わせて，言語学的研究の経験的基盤を築くための新しい検索ソフトウェア（一体型のコーパスプラットフォーム）の開発も進められる→**COSMAS**（**Co**rpus **S**earch, **Ma**nagement and **A**nalysis **S**ystem）

コーパス検索ソフトウェアの開発

- **COSMAS 1**（1991年～2003年）
- プロジェクト当初に，労力を要する検索に関する業務負担を軽減するため，独立した新規のプロジェクトとして進める決断を行う

コーパス検索ソフトウェアの開発（続き）

- **COSMAS 2**（1994年～）
- EU のMECOLB (1994年2月から1995年8月まで) による助成
- 連邦国防省のSERGES（1995年10月から1999年3月まで）による助成
- 1995年以降，COSMAS1と2は組織上別のプロジェクトとされる
- COSMAS 2の初期のプロトタイプはデンマークおよびアメリカで製造される
- 2002年：windows版がリリースされる
- 2005年：web版，server版がリリースされる
- 2006年：script版がリリースされる
- 主な検索機能：コンコーダンス検索，コロケーション検索

コーパス検索ソフトウェアの開発（続き）

- しかし，COSMAS 2では現在の複雑で多様なニーズに対応することが困難であることからその後継として2011年より**KorAp**（**Kor**pus**a**nalyse**p**lattform）が開発される
- KorApの特徴：
  - →COSMAS 1/2の機能の継承
  - →データ処理の高速化
  - →正規表現の利用
  - →複数の検索クエリー言語の利用
  - →情報付与：トークン（単語など），スパン（文・フレーズなど），関係（依存関係など），属性（HTML要素の属性など）

ドイツのコーパス

【書き言葉コーパス】

- DeReKo （IDS）
- DWDS（BBAW）

DeReKo

- **DeReKo**=Das **De**utsche **Re**ferenz**ko**rpus（ドイツ語書き言葉参照コーパス）
- 総語数：約576億語（2024年1月9日時点，IDS公式HP）
- 目的：言語研究の経験的基盤
- ジャンル：（膨大な数の）新聞記事，文学テキスト，学術テキスト，一般向けの科学的なテキストをはじめとする多様なテキストが含まれている
- 年代：近世～

DeReKo（続き）

- 費用：商用目的でない限り．COSMAS 2およびKorApを介して，無償で利用できる
- 更新状況：毎年データが追加更新される
- 助成：バーデン・ビュルテンベルク州（1990年5月～2002年3月まで）
- 問題点（発表者の個人的印象）：
  →巨大なため中身が見えず注意も必要
  →通信社から記事を購入している新聞社も多く，文言が一致する記事が複数見られることもある
  →一定期間同一の新聞に広告を掲載していることもあり，時に分析結果に反映される

DWDSコーパス

- **DWDS**=**D**igitales **W**örterbuch der **d**eutschen **S**prache（ドイツ語デジタル辞書）
- 開発機関：ベルリン・ブランデンブルク科学アカデミー（BBAW）
- 当初の目的：1952年～1977年にドイツ科学アカデミー（1972年からはドイツ民主共和国科学アカデミー）により編纂された『現代ドイツ語辞典（Wörterbuchs der deutschen Gegenwartssprache）』の電子化
- 主な目的：**DWDS-Kernkorpus**（DWDS中核コーパス）を作成し，20世紀（後に21世紀も）の一言語辞書の経験的基盤として役立つこと
- DWDSの三つの要件：
  →大規模であること
  →ジャンルのバランスがとれていること
  →重要な文学作品や作家を数多く含むこと

DWDS-Kernkorpusの開発

- この要件を踏まえ，時代と年代において均衡性が担保されたDWDS-Kernkorpusを作成（その際，BNCが意識される）
- 総語数：約1億語
- 年代：1900年～1999年
- ジャンル：文学テキスト（26%），新聞（27%），学術テキスト（22%），ノンフィクション（20%），話し言葉（5%）※ただしコーパスの利用にあたって話し言葉は除外される

DWDSコーパス（続き）

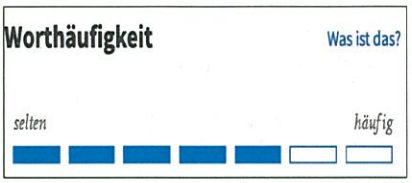
- 総語数：約705億語（DWDS公式HP）
- 現在のカテゴリー：5カテゴリー
  1）メタコーパス：歴史コーパスなど
  2）参照コーパス：DWDS-Kernkorpus(1900–1999)，DWDS-Kernkorpus 21 (2000–2010)，DTA-Kernkorpus (1598–1913)
  3）新聞コーパス：Frankfurter Allgemeine Zeitung紙など
  4）Webコーパス：IT関連ブログコーパスなど
  5）特殊コーパス：話し言葉コーパス，映画字幕コーパス，旧東ドイツコーパス，連邦議会コーパス，チャットコーパス，Wikipediaコーパスなど

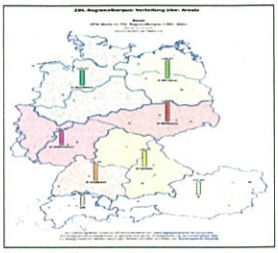
DWDSコーパス（続き）

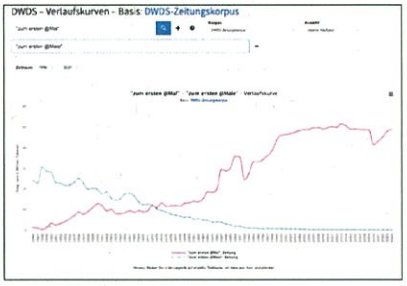
- 助成：DWDS＝ドイツ科学アカデミー連合の一部，ドイツ語デジタル辞書学センター（ZDL）の一部でもある。前者は連邦政府・州政府より，後者は連邦教育研究省（BMBF）より助成を受けている

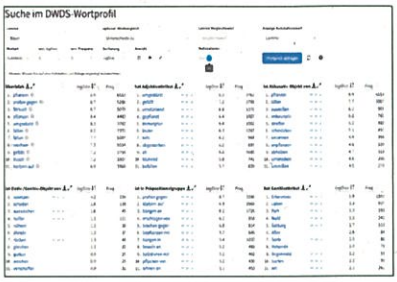
DWDSコーパスの特徴

- 視覚化されて使いやすくさまざまな機能がある
- 利用可能な項目
  - →辞書
  - →頻度情報
  - →地理的分布
  - →時系列変化
  - →コンコーダンス
  - →コロケーション
  - →類義語比較









オーストリアのコーパス

【書き言葉コーパス】

- Austrian Media Corpus (amc)

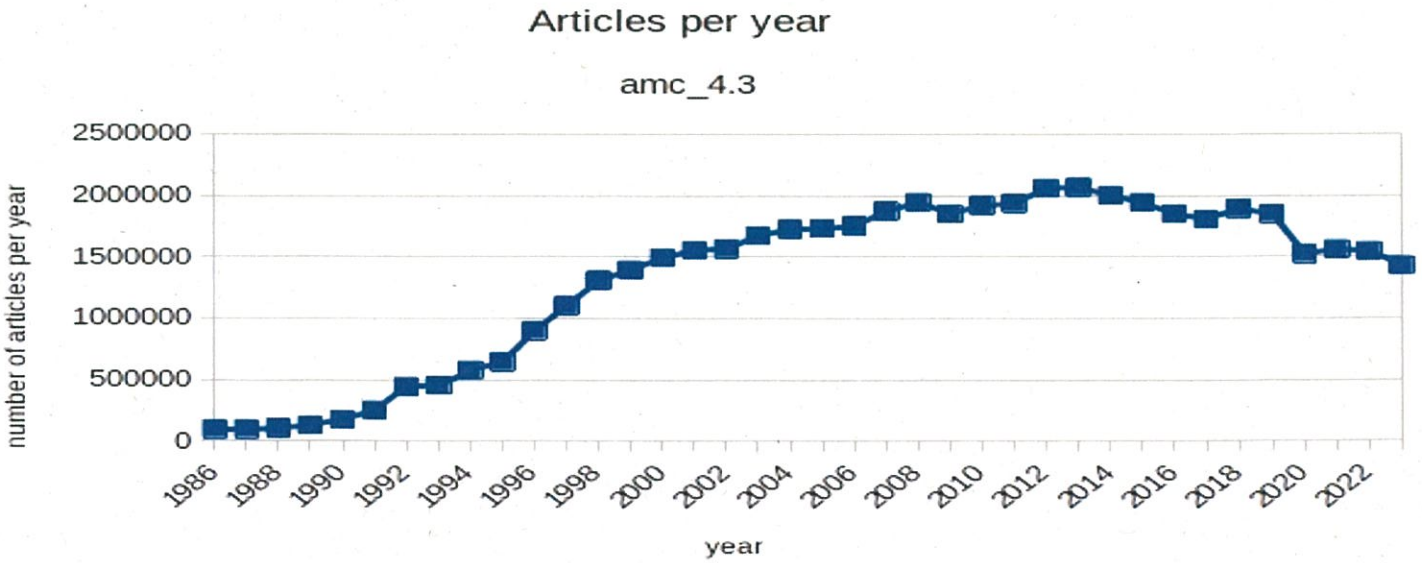
【話し言葉コーパス】

- DiÖ-Korpus
- WBÖ-Korpus

Austrian Media Corpus (amc)

- ●書き言葉コーパス
- ●オーストリアデジタルヒューマニティー・文化遺産センター（ACDH-CHオーストリア科学アカデミー (ÖAW)の一組織）
- ●テキストサイズ：約127億語（Version: 4.24q3）
- ●助成：オーストリア通信社 (APA)
  →資金およびデータの提供を受けている
- ●記事数：5100万件
- ●年代：1986年～
- ●更新状況：毎年データが追加更新される

amcの収録記事数の推移（公式HPに基づく）



※ 1990年以降記事数が急速に増加し，2000年以降は毎年150万以上の記事が収録されている

出典：https://amc.acdh.oeaw.ac.at/dokumentation/amc-in-zahlen/

Austrian Media Corpus (amc)（続き）

- テキストジャンル：ジャーナリズム散文（紙媒体　81%，通信社　17%，テレビ　2%，ラジオ　1%未満）
- 収集地域：オーストリア全土
- 情報付与：テキスト情報（メディア名，メディアタイプ，出版年，地域名など），品詞情報，固有表現（Named Entitiy）
- 利用方法：Sketch Engine，NoSketch Engineを介して利用可能
- 公式HP（URL）：https://amc.acdh.oeaw.ac.at/

DiÖ-Korpus

- DiÖ：**D**eutsch **i**n **Ö**sterreich（オーストリアにおけるドイツ語）
- 話し言葉コーパス
- ウィーン大学，ザルツブルク大学，グラーツ大学，オーストリア科学アカデミー（ÖAW）により共同運営される
- 助成：連邦教育科学研究省（BMBWF）のオーストリア科学基金（FWF）の助成
- 2016年に開始

DiÖ-Korpus（続き）

- データの内容：インタビュー，学校でのグループディスカッション，友人間の会話，制御された音声データ（翻訳を含むテキストの朗読など），言語産出実験，アンケートデータ
- 地域：56地点
- 収録人数：約470名（都市の実態を踏まえ，L2話者も含まれる。どの程度かは不明）
- 規模：約1057時間
- 利用状況：非公開

WBÖ-Korpus

- WBÖ：**W**örterbuch der historischen **b**airischen Mundarten in **Ö**sterreich und Südtirol（オーストリアと南チロルにおける歴史的バイエルン方言辞典）のプロジェクトの一つ
- WBÖ は1911年以降のバイエルン方言の語彙を記録する方言辞典。1963年に出版が始まり，5巻（Eまで）が紙媒体で刊行されている。 Fからは2020年よりオーストリア語彙情報システム（LIÖ）のサイト上で一般公開されている
  →URL：https://lioe.dioe.at/
- オンライン化に伴い構築されたコーパス

WBÖ-Korpus

- データの内容：手書きメモ（約85％はWBÖの作成時に収集されたD以降のメモ）
- 助成：オーストリア科学アカデミー（ÖAW）
- プロジェクト期間：2016年11月～2038年12月
- 規模：公式サイトでは記載なし
- 利用状況：非公開

2 コーパスの活用例（ドイツ）

IDSが想定するDeReKoの応用分野

- 国内のみならず海外でのドイツ語研究
- 心理学，神経学，認知科学，言語療法，コミュニケーション科学，メディア学，統計学
- 活用例・実践例の収集と公開

→活用例・実践例のリスト（URL）：

https://www.ids-mannheim.de/digspra/kl/projekte/korpora/bib/

- リストの大半はドイツ語学論文（学位論文を含む）と書籍
- その他，辞書（ドイツ：LEOs Wörterbücher，ハンガリー：LINGEA，日本：『アクセス独和辞典』三修社），授業資料などがある

DeReKoの活用例

- ドイツ語正書法委員会の事務局が設置されるIDSでは，国民的議論にも発展したドイツ語の正書法において中心的役割を担っており，分析にDeReKoの一部を提供している

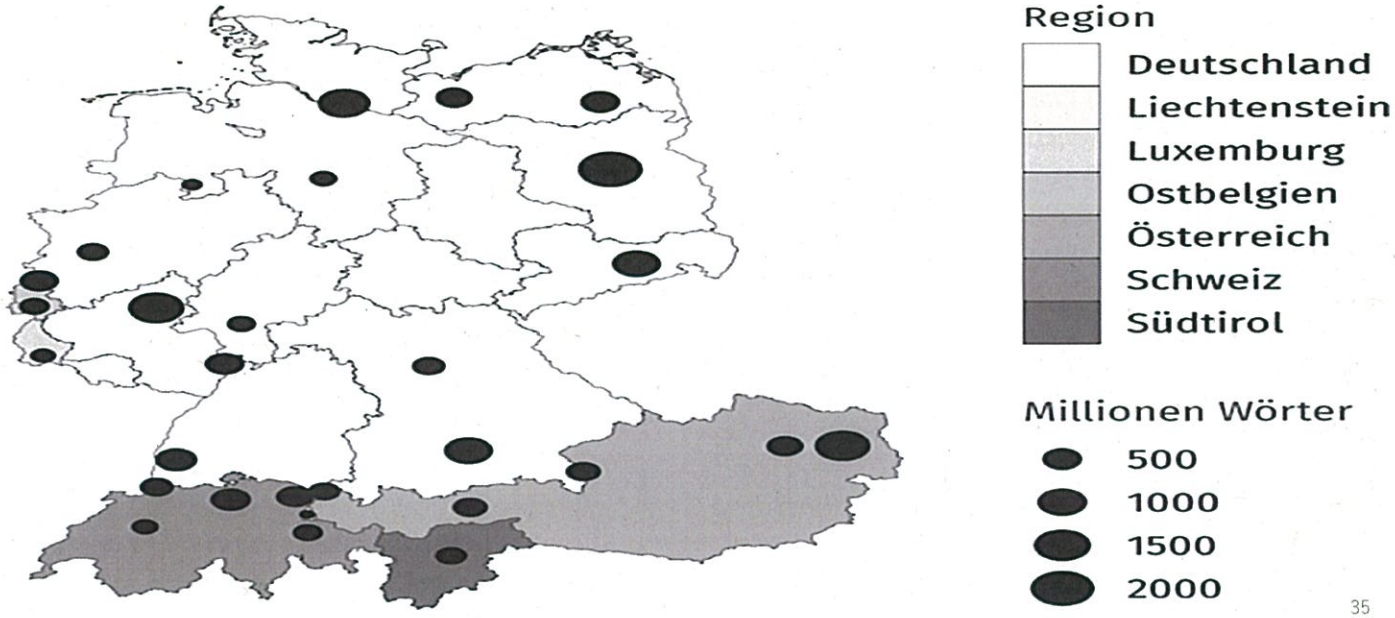
正書法（Rechtschreibung）とは

- 20世紀初頭から続く綴りに関する規則
- さまざまな綴り方が見られ，学校や官公庁では統一的な規則が必要とされたことに遡る
- 1996年のウィーン基本合意書で，正書法の大幅な改正を行うことがドイツ語圏8カ国で取り決められる→1998年8月1日より施行。2005年7月31日までが移行期間
- しかし，大幅改正への批判が大きいことから，2004年にドイツ語正書法委員会（Rat für deutsche Rechtschreibung，7カ国（ただしルクセンブルクはオブザーバー参加），41名，事務局はIDS）が設置され，2006年8月1日より修正が施される

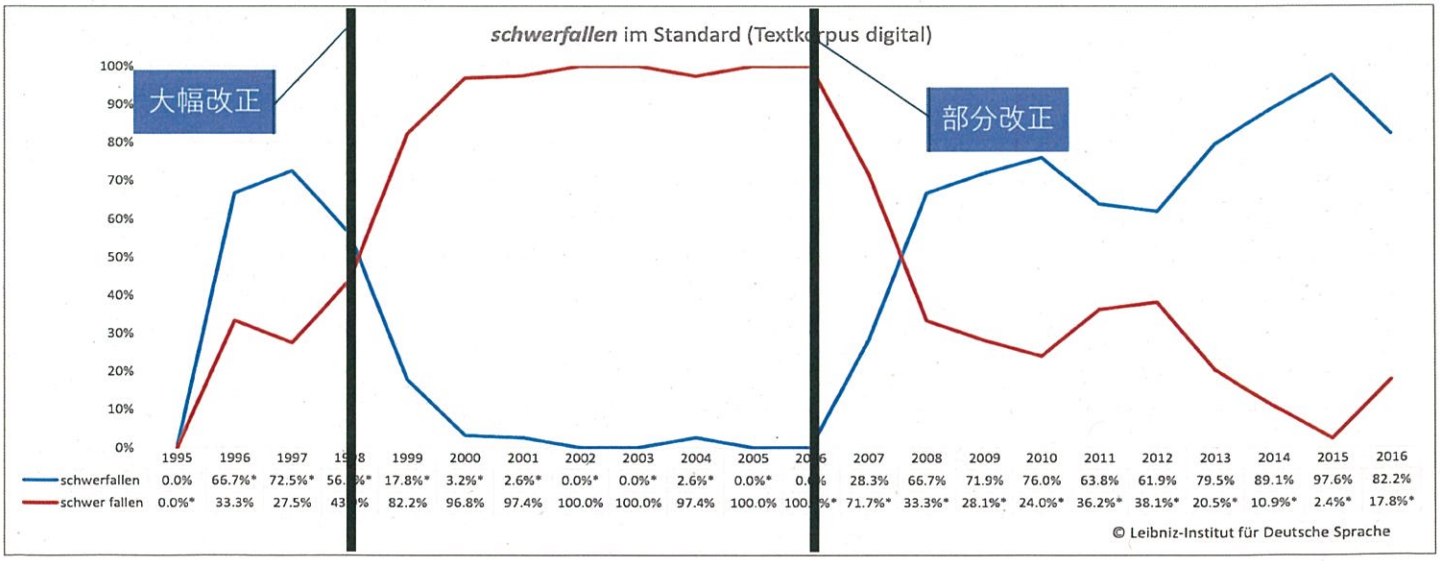
コーパスによる綴り方の観測

- ドイツ語正書法委員会では，設立当初より綴り方の実態を捉えるためにコーパス（書き言葉）を活用している
- コーパスは委員会の構成員である辞書出版社DudeおよびWahrig（現在は廃刊）の辞書編纂用のコーパスと，DeReKoおよびamcを土台として，委員会構成員のすべて地域をカバーし，ジャンルのバランスを考慮して正書法用に構築されている
- 2017年以降，新たな公式の語彙目録の作成のために構築された現在のOrthografische Kernkorpus（OKK，正書法中核コーパス）は，140億語以上の総語数を有し，正書法の大幅改正が取り決められる1年前の1995年から現在までのドイツ語圏の書き言葉を収録している（Rat für deutsche Rechtschreibung, 2024）
- OKKはIDSのコーパス検索システムの一つであるKorAPを通して利用できる
- なお，コーパスは委員会設立以前から用いられており，当初はDudenおよびWahrig が辞書編纂用に構築したコーパスが重要な役割を果たしたとされる（Krome 2022:28）

OKKの国・地域とコーパスサイズ
Diewald et al. 2024:333



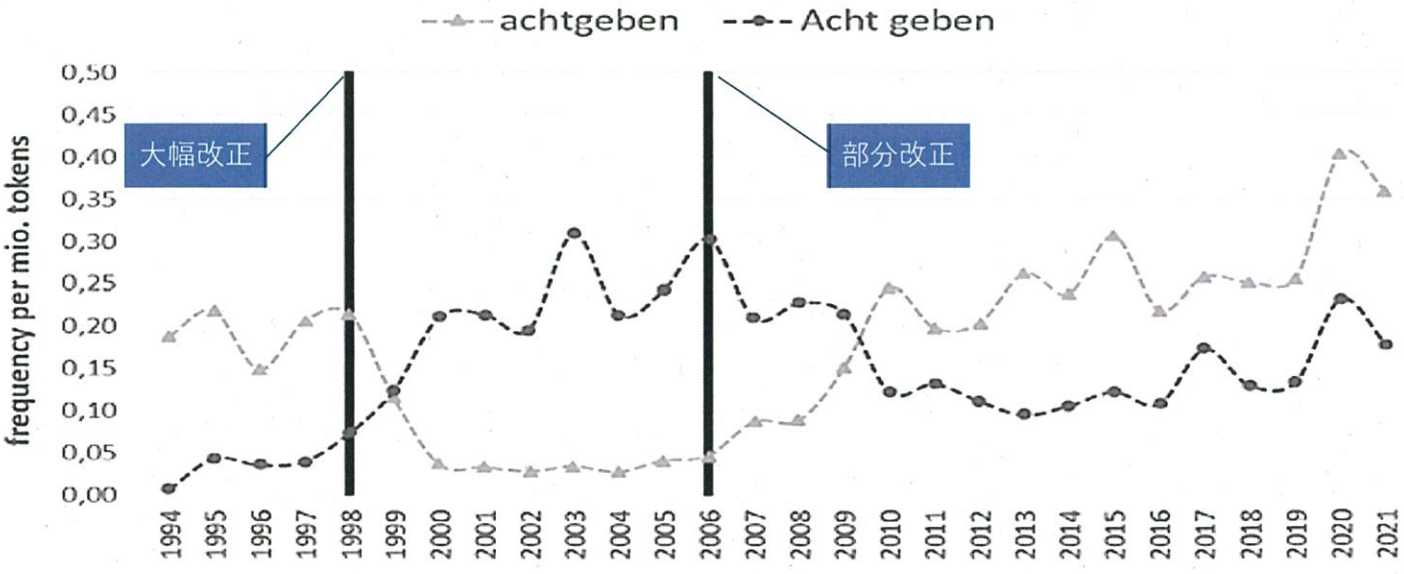
【OKKの活用例：正書法】 Krome 2022:29
綴り方の変化： schwerfallenとschwer fallen（難しい）



※ 最初の改革後に変化が見られるが，その後の部分改正により，また旧形式に戻る。
ただし，反義語のleichtfallen（容易である）は同様の傾向を強く示さない

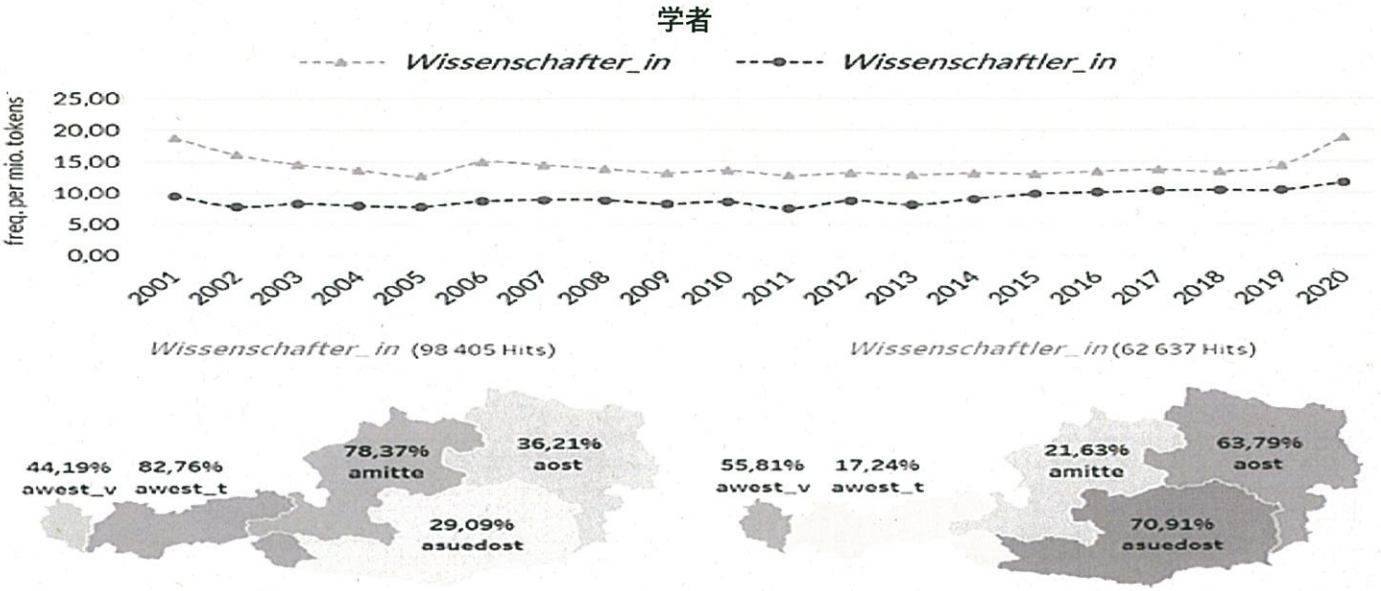
2 コーパスの活用例（オーストリア）

【amcの活用例：正書法】 Dorn et al. 2022:49
綴り方の変化：achtgebenとAcht geben（気をつける）



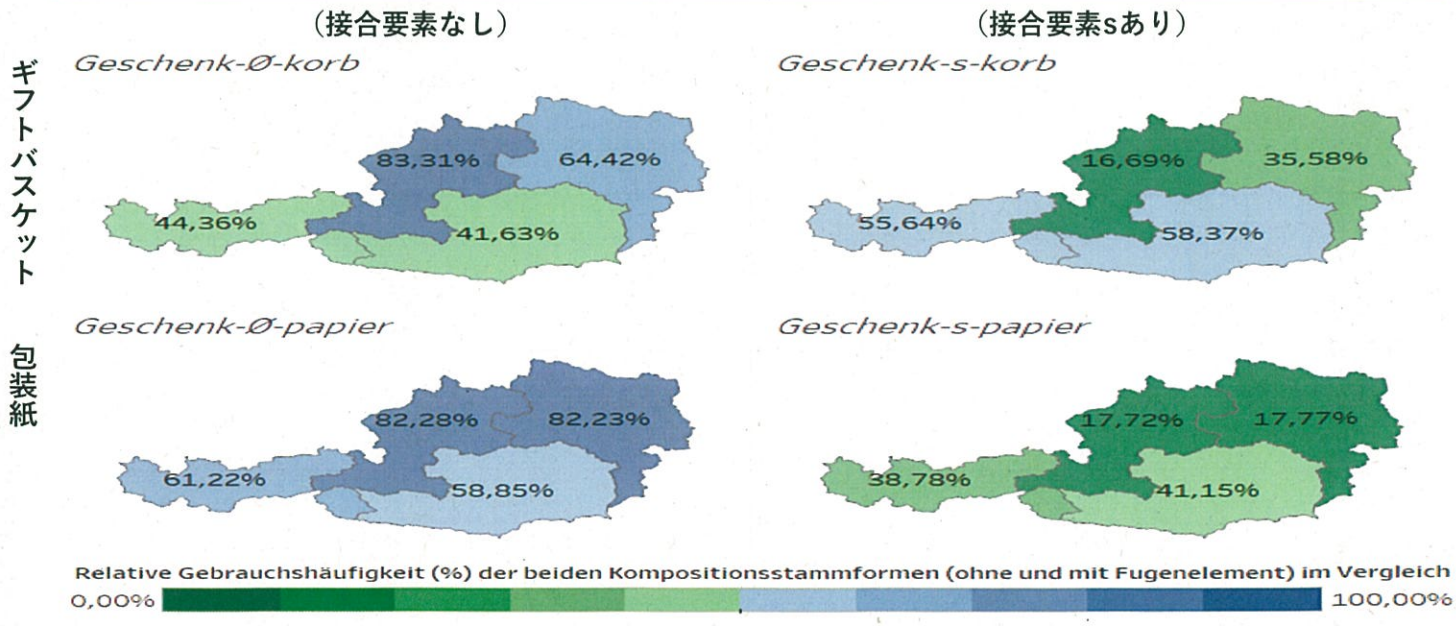
※ 最初の改革後に変化が見られるが，その後の部分改正により，また旧形式に戻る

【amcの活用例：地域差】 Dorn et al. 2022:52
人を表す接尾辞（-er, -ler）の分布



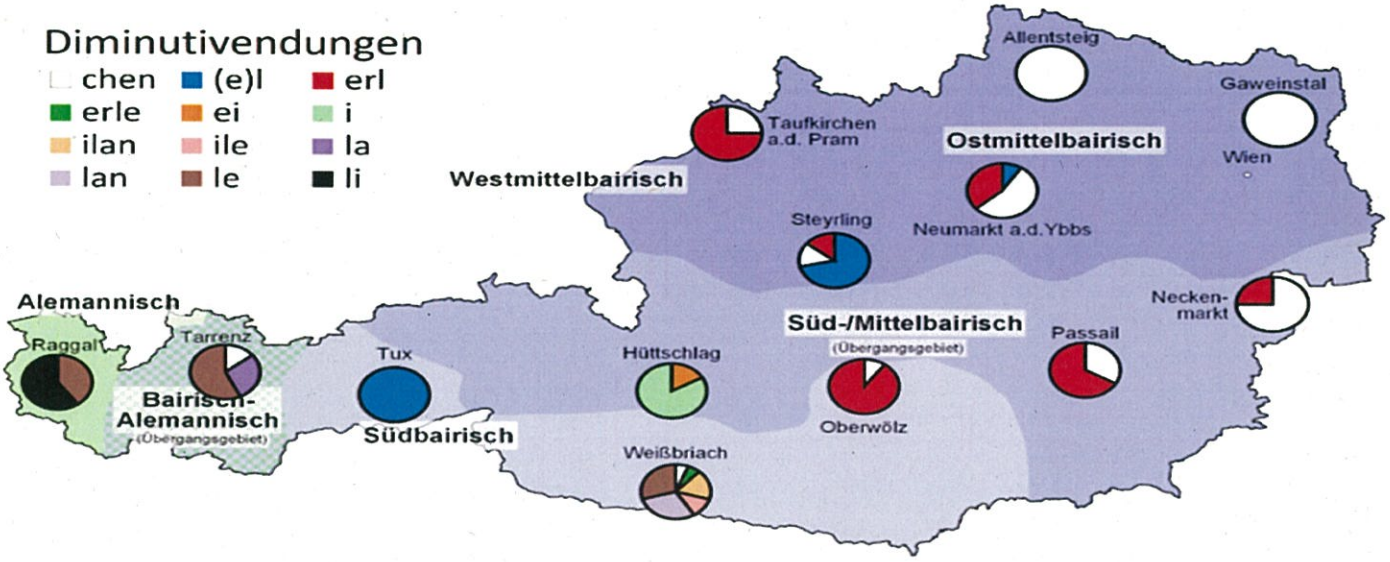
※ オーストリア全体では-erが多いが，東部，南東部では-lerが優先的に使用される

【amcの活用例：地域差】 Ziegler 2021:73
接合要素sの分布



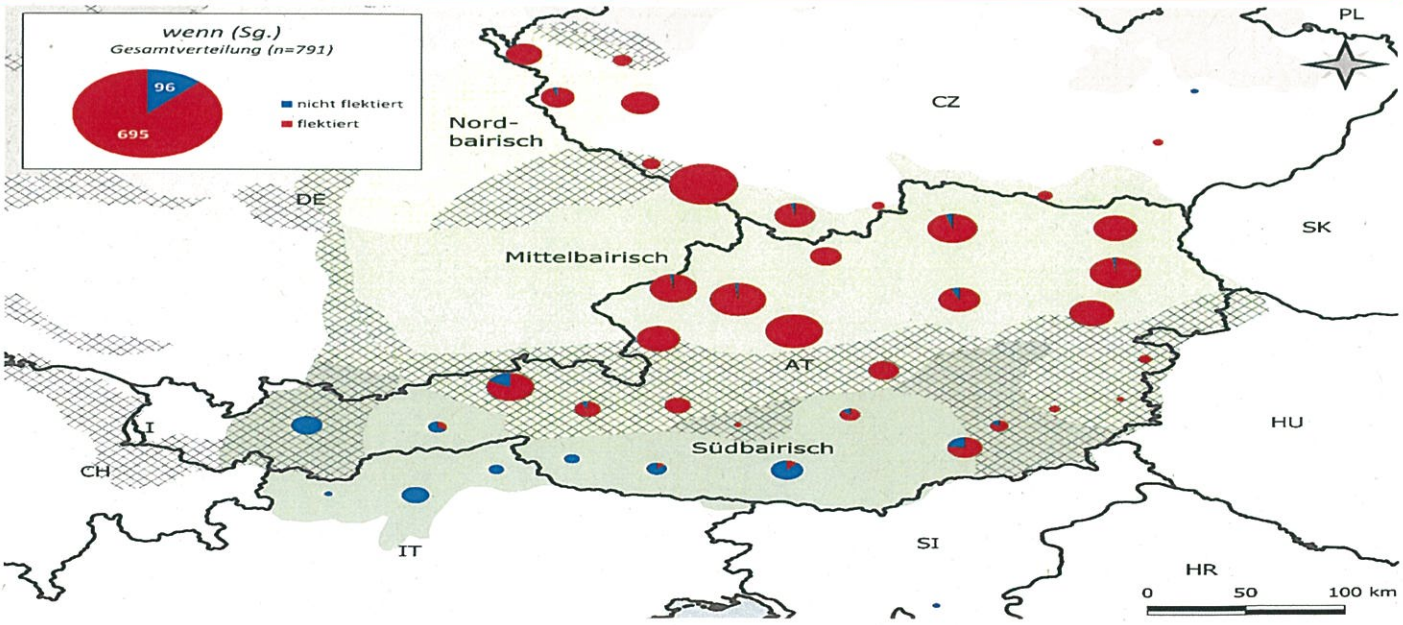
※ 接合要素の有無は地域によっても，単語によっても異なる

【DiÖの活用例：地域差】 Korecky-Kröll, 2021
縮小語尾の分布



※ 地方によって使用される縮小語尾は大きく異なる

【WBÖ（話し言葉）の活用例】Stöckle & Hemetsberger & Stütz 2021:595
wenn （英if）文中での主語（2人称単数）と動詞の一致



※南バイエル方言（Südbairisch）を中心に主語と動詞の一致がみられない地域がある[42]

まとめ

- ドイツ語の主要コーパスは，大規模なものが多く，商用目的でない限り，とくにドイツでは無償利用が基本とされる
- このことは，公的資金による継続的な支援によって成り立っている
- 助成額については，公式HPでは公表されていない
- オーストリアでは民間の外部資金（通信社）によるプロジェクトもある（→amc）

まとめ（続き）

- ドイツ語の主要コーパスはドイツ語研究や教育への応用を主たる目的としており，ドイツ語研究にコーパスを利用するケースが大半を占める（→ただし，国はその重要性を認め，長期間にわたり公的資金を投じている）
- オーストリアでは，ドイツ以上に地域特性や方言（または言語地理学）に関する研究も多く，コーパスの利用目的がドイツとは異なる部分もある
- ドイツ語コーパスの社会貢献としては，ドイツ語正書法委員会においての貢献（綴り方の観測）が挙げられる

引用文献

- Diewald, N. et al. (2024). Das Orthografische Kernkorpus (OKK) in DeReKo: Zusammensetzung, Analyse- und Zugriffsmöglichkeiten über KorAP, In Krome, S. (Hrsg.) *Orthographie in Wissenschaft und Gesellschaft: Schriftsystem – Norm – Schreibgebrauch* (PP.329-343). De Gruyter.
- Dorn, A. et al. (2022). Die österreichische Presselandschaft digital: Das Austrian Media Corpus (amc) und sein Potenzial für die Linguistik, In Kupietz, M. & Schmidt, T. (Hrsg.) *Neue Entwicklungen in der Korpuslandschaft der Germanistik* (Beiträge zur IDS-Methodenmesse 2022) (pp. 43-55). Narr.
- Geyken, A. (2007). The DWDS Corpus: A reference corpus for the German language of the 20th century. In Fellbaum, C. (Ed.) *Collocations and Idioms* (pp. 23-42)..Continuum Press.
- Korecky-Kröll, K. (2021). Wo in Österreich werden beim Einschlafen "Schäfchen", "Schafilan" oder "Lampl" gezählt?. *DiÖ-Online*. https://www.dioe.at/artikel/2839
- Krome, S. (2022). Der Rat für deutsche Rechtschreibung im Spiegel von Sprach-, Norm- und Schreibwandel. Bilanz einer wechselvollen Geschichte, *Sprachreport, 38*(1), Leibniz-Institut für Deutsche Sprache, 26-37.
- Lenz, A. N. (2023). Korpora zur deutschen Sprache in Österreich, In Deppermann, A. et al. (Hrsg.) *Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial* (pp. 53-69). De Gruyter.
- Rat für deutsche Rechtschreibung. (2024). Amtliche deutsche Rechtschreibung: Überarbeitetes Regelwerk und Neufassung Wörterverzeichnis für Schule und Verwaltung verbindlich, *Pressemitteilung* (03.07.2024). https://www.rechtschreibrat.com/DOX/RfdR_PM_2024-07-03_Aktualisierung_Regelwerk.pdf
- Stöckle, P. & Hemetsberger, C. & Stütz, M. (2021). Die WBÖ-Belegdatenbank als Quelle für syntaktische Analysen – Möglichkeiten, Grenzen, Perspektiven, *Wiener Linguistische Gazette, 89*, 579–626.
- Teubert, W. & Belica, C. (2014). Von der linguistischen Datenverarbeitung am IDS zur "Mannheimer Schule der Korpuslinguistik", *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, Institut für Deutsche Sprache, 298-319.
- Zieger, T. (2021). Über Geschenk-s-körbe und Schokolade-Ø-torten (zu runden Geburtstagen), *Wiener Linguistische Gazette, 89*, 55–83.

ご清聴ありがとうございました

Vielen Dank für Ihre Aufmerksamkeit!