言語資源小委(R8.6.2)

資 料 4

ドイツ語コーパス(DeReCo)拡充の背景

●ドイツ語コーパスの拡充

2025 年にマンハイムのドイツ語研究所:ドイツ語コーパス(ドイツ語書き言葉参照コーパス DeReKo;Das Deutsche Referenzkorpus)の大幅な拡充があり、2026 年1月時点で 638 億語規模のコーパスとなった。(参考:2024 年 1 月時点 約 576 億語)

●拡充の背景

・著作権法などの法律の改正などは行われていないので、直接的な要因とは考え難い。

・拡充データとしては、Wikipedia ユーザーによるオンライン上でのディスカッションのデータが非常に大きく、拡充量としてはこれを使えたことの影響が強い。

＜ドイツ語研究所の更新情報(英語)＞

https://jpn01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.ids-mannheim.de%2Fen%2Fdigspra%2Fpb-s1%2Fprojects%2Fcorpus-development%2F&data=05%7C02%7Cjinsuzu%40mext.go.jp%7C5f70ee50bddd4c7c51b308dea84a1a21%7C545810b036cb4290892648dbc0f9e92f%7C0%7C0%7C639133233384926877%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=0Byoznp%2Bn3Yf%2B3gdwp%2BeA%2FaZ1%2FMzUfZHfN0vpYkv4bY%3D&reserved=0

・政府主導の大規模プロジェクト(「国家研究データ基盤」:2020年～)の一環として、ドイツ国内の分散するドイツ語に関するデジタルデータを集約し、研究利用できる基盤を構築する「言語・テキストデータ基盤構築プロジェクト(Text+)」が始まったことで、潤沢な予算が確保できるようになったことが強く影響している。

Text+における分担　(幹事)ドイツ語研究所:書き言葉コーパス
テュービンゲン大学:話し言葉データ
ゲッティンゲン大学図書館:古文書・歴史書
ベルリン・ブランデンブルク科学アカデミー:辞書

ドイツ語研究所の予算額の推移

年度	公的予算計	（国）	（州）	外部資金	合計
2015 年度	1121	（555	555）	1238	2359
2016 年度	1272	（650	613）	1414	2686
2017 年度	1297	（690	607）	1459	2756
2018 年度	1316	（708	598）	1488	2804
2019 年度	1340	（732	586）	1466	2806
2020 年度	1360	（759	586）	1548	2908
2021 年度	1379	（769	596）	1513	2892
2022 年度	1402	（780	607）	1715	3117
2023 年度	1496	（827	655）	1714	3210
2024 年度	1570	（854	706）	1776	3346
2025 年度	1580	（845	723）	1739	3319

（万ユーロ）