Europäische Sprachmodelle: Herausforderungen, Initiativen, Lösungsansätze

Prof. Dr. Georg Rehm, DFKI GmbH
georg.rehm@dfki.de

dfki ai
Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for
Artificial Intelligence

07.12.2023 – KI in Bibliotheken: Neue Wege mit großen Sprachmodellen? – DNB, Frankfurt

Eine (sehr) kurze Geschichte der Erschließung – In drei Phasen

- Früher: Karteikarten erschließen Metadaten über Werke
- Dann: Einzug der digitalen Kataloge, die aber ebenfalls „nur“ Metadaten erschließen
- Kurze Zeit später: Suchmaschinen (Plural) erschließen eine heterogene Mischung aus Metadaten und Volltexten in diversen analogen und digitalen Bibliotheken sowie Katalogen

DEUTSCHE NATIONAL BIBLIOTHEK

English Gebärdensprache Leichte Sprache Hilfe und Kontakt

MENÜ BENUTZUNG DNB PROFESSIONELL

FACHVERANSTALTUNG

KI IN BIBLIOTHEKEN: NEUE WEGE MIT GROSSEN SPRACHMODELLEN?

Beginn 07.12.2023, 13:00 Uhr
Ende 08.12.2023, 13:00 Uhr
Wo Frankfurt am Main

KI in Bibliotheken – Eine Vision

- Große Sprachmodelle werden auch in Bibliotheken Einzug halten
- Bereits jetzt verbessern sie Basisfunktionen: OCR, OLR, Search, Discovery etc.
- In einigen Jahren dann die Revolution:

**Mittels ScienceGPT mit der wissenschaftlichen Literatur sprechen,
mit allen wissenschaftlichen Publikationen in den Dialog treten,
diese aktiv konsultieren, Forschung gemeinsam mit den Quellen betreiben!**

- Basis: Ein LLM wird auf Basis möglichst *aller* jemals erschienenen wissenschaftlichen Veröffentlichungen in allen Sprachen trainiert und mit Chat-Funktionalität ausgestattet.
- Voraussetzungen (Auswahl): Die gesamte wissenschaftliche Fachliteratur liegt digital vor (in allen Sprachen); die Rechteproblematik ist geklärt; LLMs halluzinieren nicht mehr; Faktualität liegt bei 100%; LLMs beherrschen Mathematik und alle Naturgesetze (durch das Training); die Ausgaben von LLMs liefern Quellenangaben; strukturiertes Wissen ist im LLM nutzbar.

Überblick

- Einleitung
- Modelle und digitale Sprachgerechtigkeit
- Daten
- Wissenschaftliche Daten
- Schlussfolgerungen

Modelle

Kontext: Große Sprachmodelle – Large Language Models (LLMs)

- LLMs sorgen in allen Domänen für massive Disruptionen (GPT-3, ChatGPT, Gemini etc.)
- Sie stellen den bedeutendsten Durchbruch in der KI-Forschung der jüngsten Vergangenheit dar
- LLMs werden mit immens großen Datenmengen trainiert (Sprachdaten, d.h. Text)
- Sie nutzen Hunderte von Terabytes, oft Petabytes (Billionen Tokens), an Sprachdaten sowie auch Bild-, Video- und Audiodaten
- Besonderheit: Für fast alle Sprachen Europas existieren Herausforderungen

The Washington Post
Democracy Dies in Darkness

Sign in

BUSINESS

ChatGPT Shows Just How Far Europe Lags in Tech

Analysis by Lionel Laurent | Bloomberg

February 21, 2023 at 2:12 a.m. EST

Comment 1 Gift Article Share

Europe is where ChatGPT gets regulated, not invented. That's something to regret. As unhinged as the initial results of the artificial-intelligence arms race may be, they're also another reminder of how far the European Union lags behind the US and China when it comes to tech.

Regelrechte Explosion des LT/NLP Markts: 439.85 Mrd. $ bis 2030

Natural Language Processing Market Size, Share & Trends Analysis Report By Component, By Deployment Model, By Enterprise Size, By Type, By Application, By End-use, By Region, And Segment Forecasts, 2023 - 2030

Report ID: GVR-4-68040-020-4 | Number of Pages: 100 | Format: Electronic (PDF)

Historical Range: 2017 - 2021 | Industry: Technology

https://www.grandviewresearch.com/industry-analysis/natural-language-processing-market-report

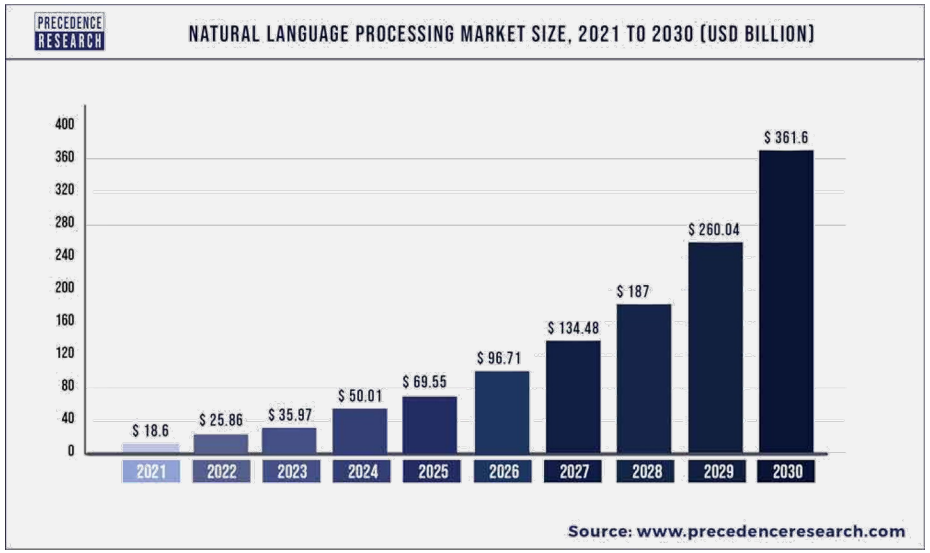
Natural Language Processing Market Report Scope

Report Attribute	Details
Market size value in 2023	USD 40.98 billion
Revenue forecast in 2030	USD 439.85 billion
Growth rate	CAGR of 40.4% from 2023 to 2030
Base year for estimation	2022
Historical data	2017 - 2021
Forecast period	2023 - 2030
Quantitative units	Revenue in USD million and CAGR from 2022 to 2030

Players leading the NLP market include-

- 3M Co. (US)
- IBM Corporation (US)
- Hewlett-Packard Co. (US)
- Oracle Corporation (US)
- Apple Inc. (US)
- Microsoft Corporation (US)
- SAS Institute Inc. (US)
- Dolbey Systems Inc. (US)
- Verint Systems Inc. (US)
- Net base Solutions Inc. (US)

US!



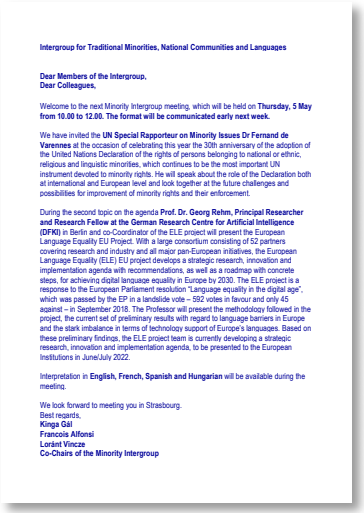
https://www.precedenceresearch.com/natural-language-processing-market

Ohne gezielte Interventionen der EU wird Europa in den kommenden Jahren weiter an den Rand gedrängt.

Europäische Initiativen

- Es existieren diverse europäische Initiativen für die Entwicklung großer Sprachmodelle
  - Forschungsprojekte in nahezu allen Ländern, z.B. Spanien, Dänemark, Slovenien, Italien etc.
  - Firmen in vielen Ländern, z.B. Finnland (Silo.ai) oder Frankreich (Mistral)
  - EU-Projekte, z.B. HPLT
  - Übergreifende europäische Initiative: ALT-EDIC – dazu später mehr
- Wichtige Initiativen in Deutschland
  - Firmen: Aleph Alpha (Heidelberg) und andere
  - Forschungsprojekte: OpenGPT-X und andere
- Herausforderungen: Datenlage speziell für europäische Sprachen; High-Performance Compute Einrichtungen; Geschwindigkeit der zentralen Big-Tech-Player vs. Geschwindigkeit Europas

Europaparlament und Europäische Kommission (2022/23)



Intergroup for Traditional Minorities, National Communities and Languages

Dear Members of the Intergroup,
Dear Colleagues,

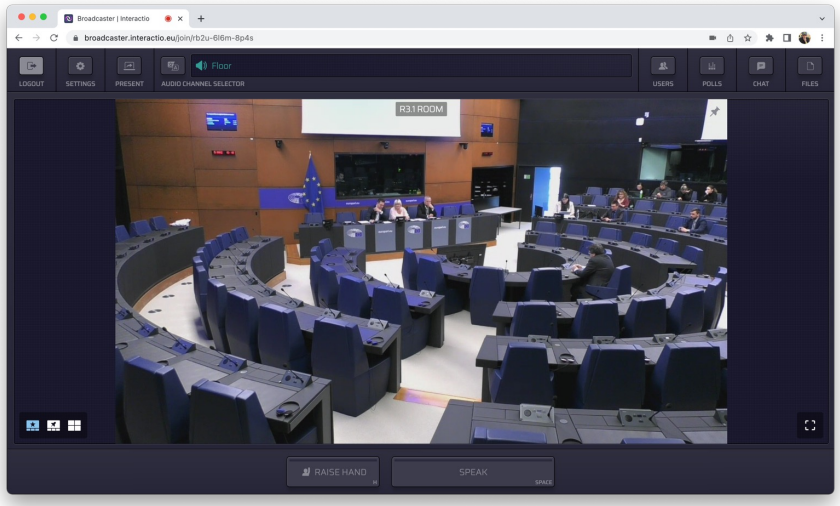
Welcome to the next Minority Intergroup meeting, which will be held on **Thursday, 5 May from 10.00 to 12.00. The format will be communicated early next week.**

We have invited the **UN Special Rapporteur on Minority Issues Dr Fernand de Varennes** at the occasion of celebrating this year the 30th anniversary of the adoption of the United Nations Declaration of the rights of persons belonging to national or ethnic, religious and linguistic minorities, which continues to be the most important UN instrument devoted to minority rights. He will speak about the role of the Declaration both at international and European level and look together at the future challenges and possibilities for improvement of minority rights and their enforcement.

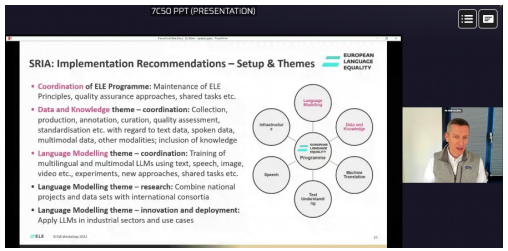
During the second topic on the agenda **Prof. Dr. Georg Rehm, Principal Researcher and Research Fellow at the German Research Centre for Artificial Intelligence (DFKI)** in Berlin and co-Coordinator of the ELE project will present the European Language Equality EU Project. With a large consortium consisting of 52 partners covering research and industry and all major pan-European initiatives, the European Language Equality (ELE) EU project develops a strategic research, innovation and implementation agenda with recommendations, as well as a roadmap with concrete steps, for achieving digital language equality in Europe by 2030. The ELE project is a response to the European Parliament resolution "Language equality in the digital age", which was passed by the EP in a landslide vote – 592 votes in favour and only 45 against – in September 2018. The Professor will present the methodology followed in the project, the current set of preliminary results with regard to language barriers in Europe and the stark imbalance in terms of technology support of Europe's languages. Based on these preliminary findings, the ELE project team is currently developing a strategic research, innovation and implementation agenda, to be presented to the European Institutions in June-July 2022.

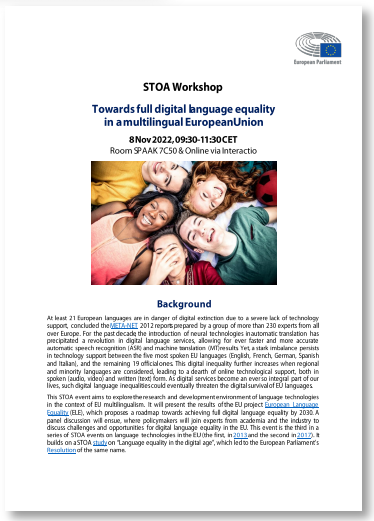
Interpretation in **English, French, Spanish and Hungarian** will be available during the meeting.

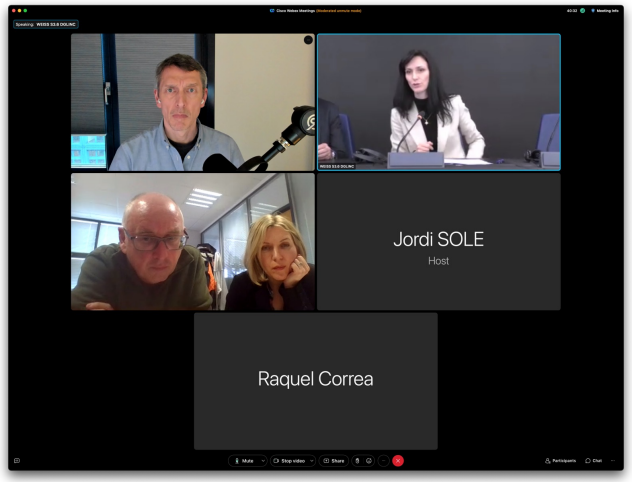
We look forward to meeting you in Strasbourg.
Best regards,
Kinga Gál
Francois Alfonsi
Loránt Vincze
Co-Chairs of the Minority Intergroup



Intergroup for Traditional Minorities, National Communities and Languages (März 2022)







Besprechung mit Commissioner Mariya Gabriel (März 2023)

STOA Workshop "Towards full digital language equality in a multilingual European Union" (November 2022)

European Language Equality (ELE) 1 und 2

META-FORUM 2022 – 8./9. Juni, Brüssel, Belgien

http://www.european-language-equality.eu

Koordinator: DCU (ADAPT Centre)

Co-Koordinator: DFKI

EUROPEAN LANGUAGE EQUALITY

Konsortium: 52 Partner (Kerngruppe: DCU – ADAPT Centre, DFKI, Charles University, ILSP, University of the Basque Country)

Ziel: *Entwicklung einer strategischen Forschungsagenda, um in Europa bis 2030 für digitale Sprachgerechtigkeit zu sorgen*

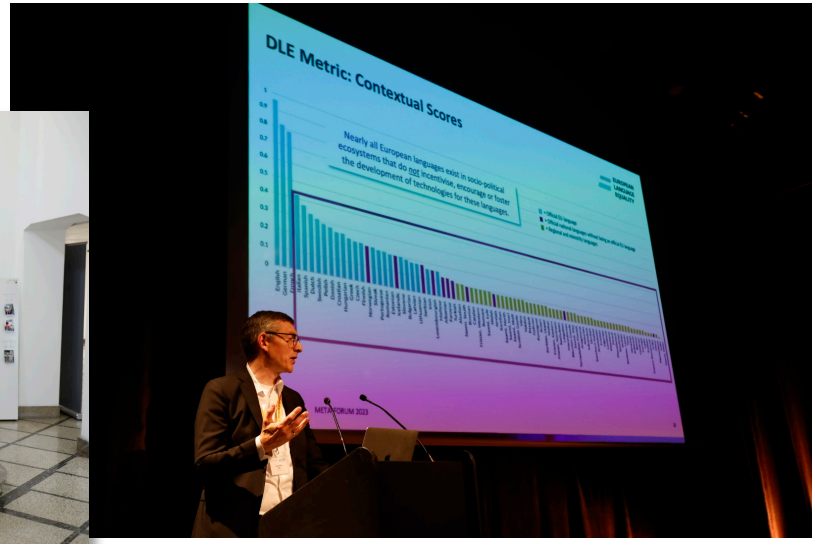
Laufzeit: 18 Monate • Januar 2021 – Juni 2022

EUROPEAN LANGUAGE EQUALITY 2

Konsortium: 7 Partner (Kerngruppe + EFNIL und ELEN)

Ziel: *Revision und Erweiterung der strategischen Forschungsagenda*

Laufzeit: 12 Monate • Juli 2022 – Juni 2023



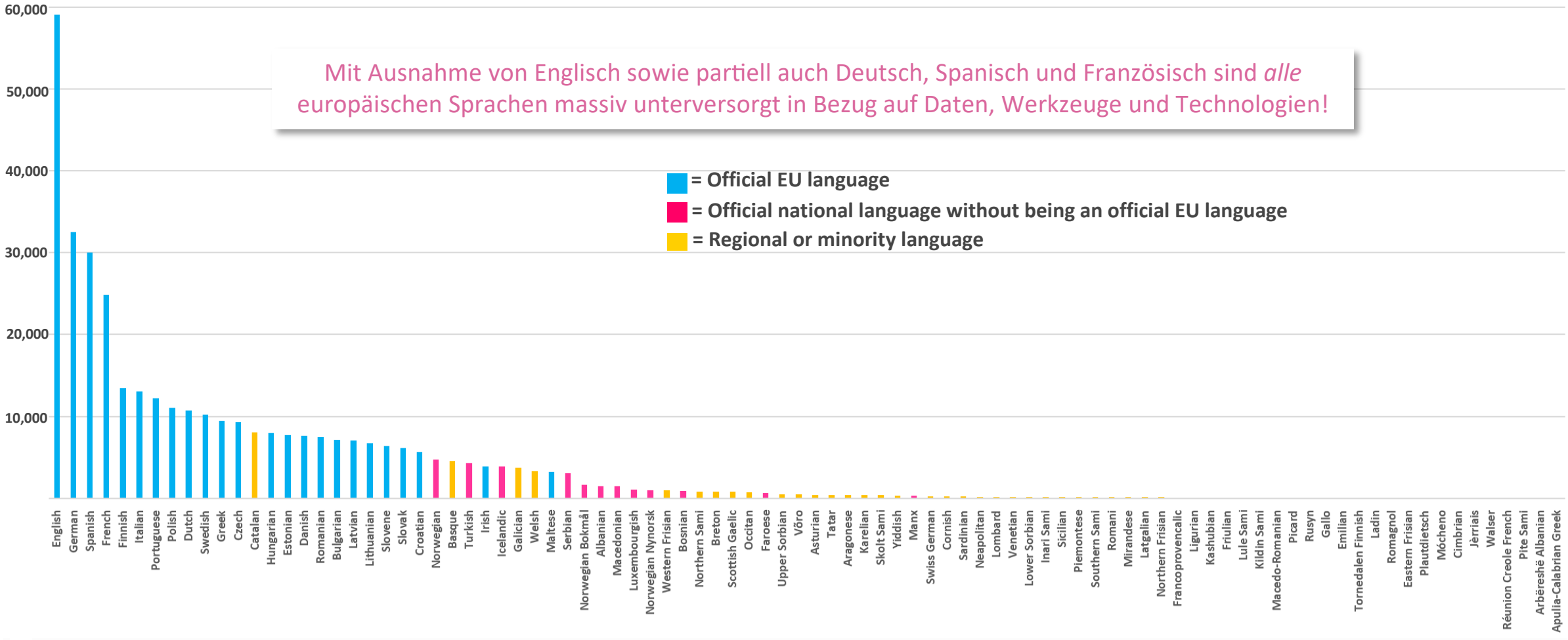
META FORUM 2023

27. Juni 2023

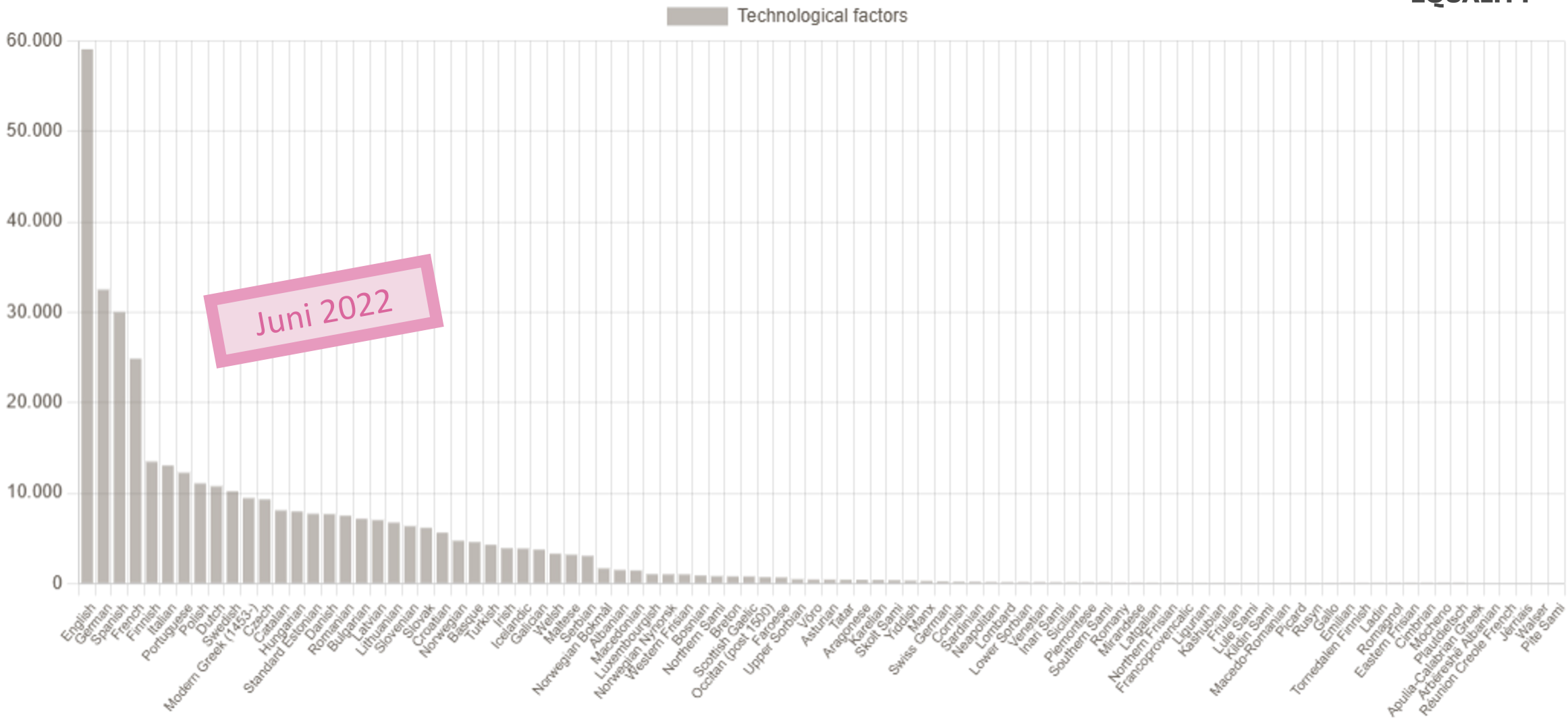
LDS

Digital Language Equality Metrik

EUROPEAN LANGUAGE EQUALITY

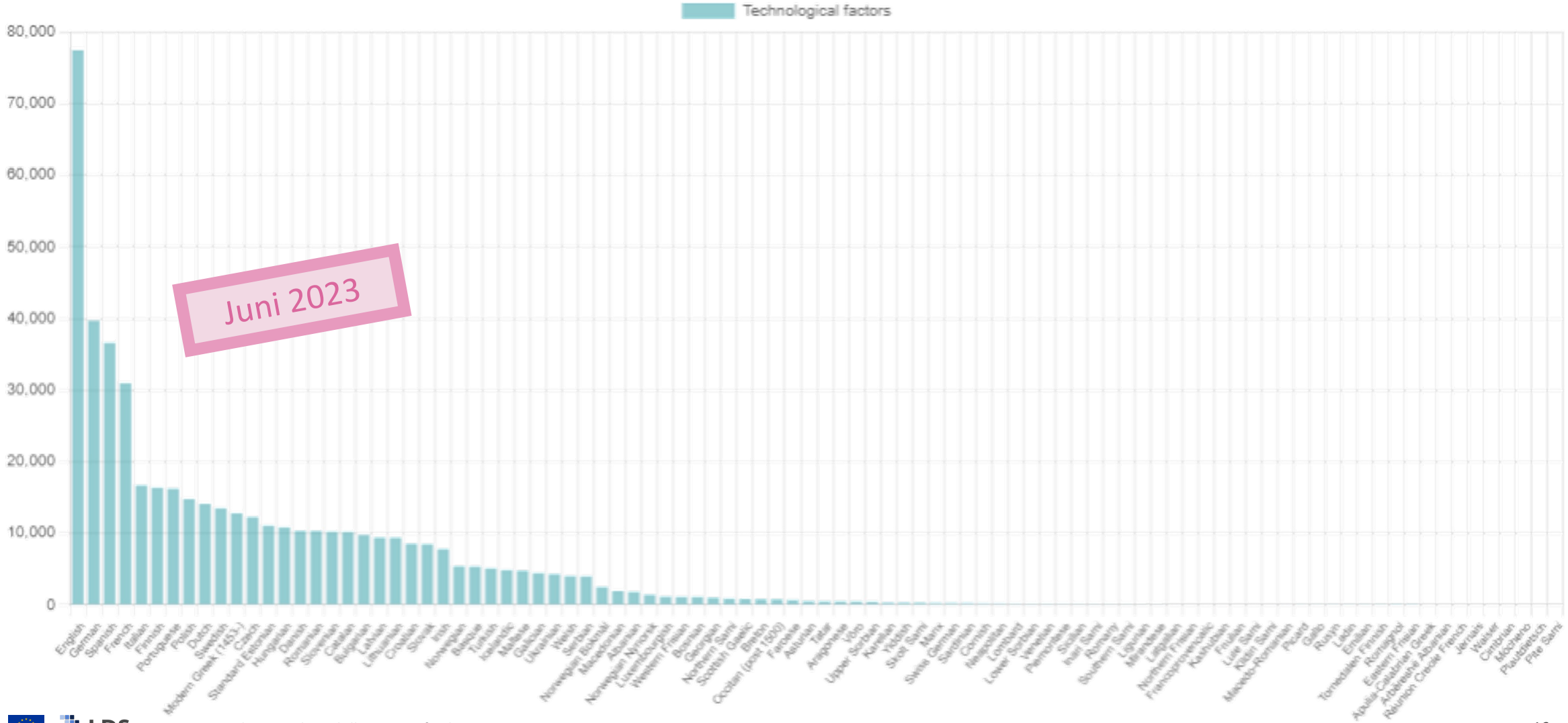


Digital Language Equality Metrik: 2022 vs. 2023 (1/3)



Digital Language Equality Metrik: 2022 vs. 2023 (2/3)

EUROPEAN LANGUAGE EQUALITY

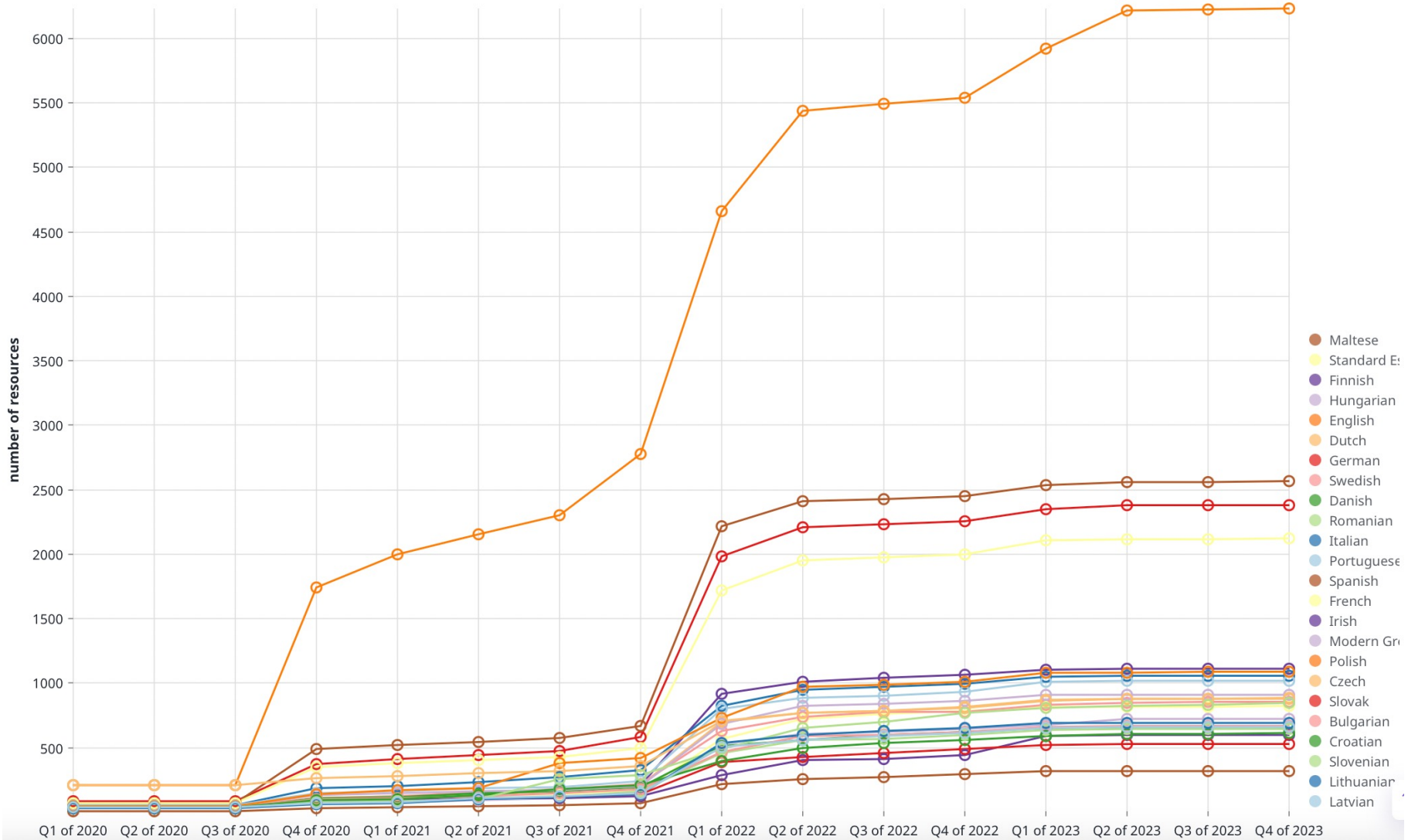


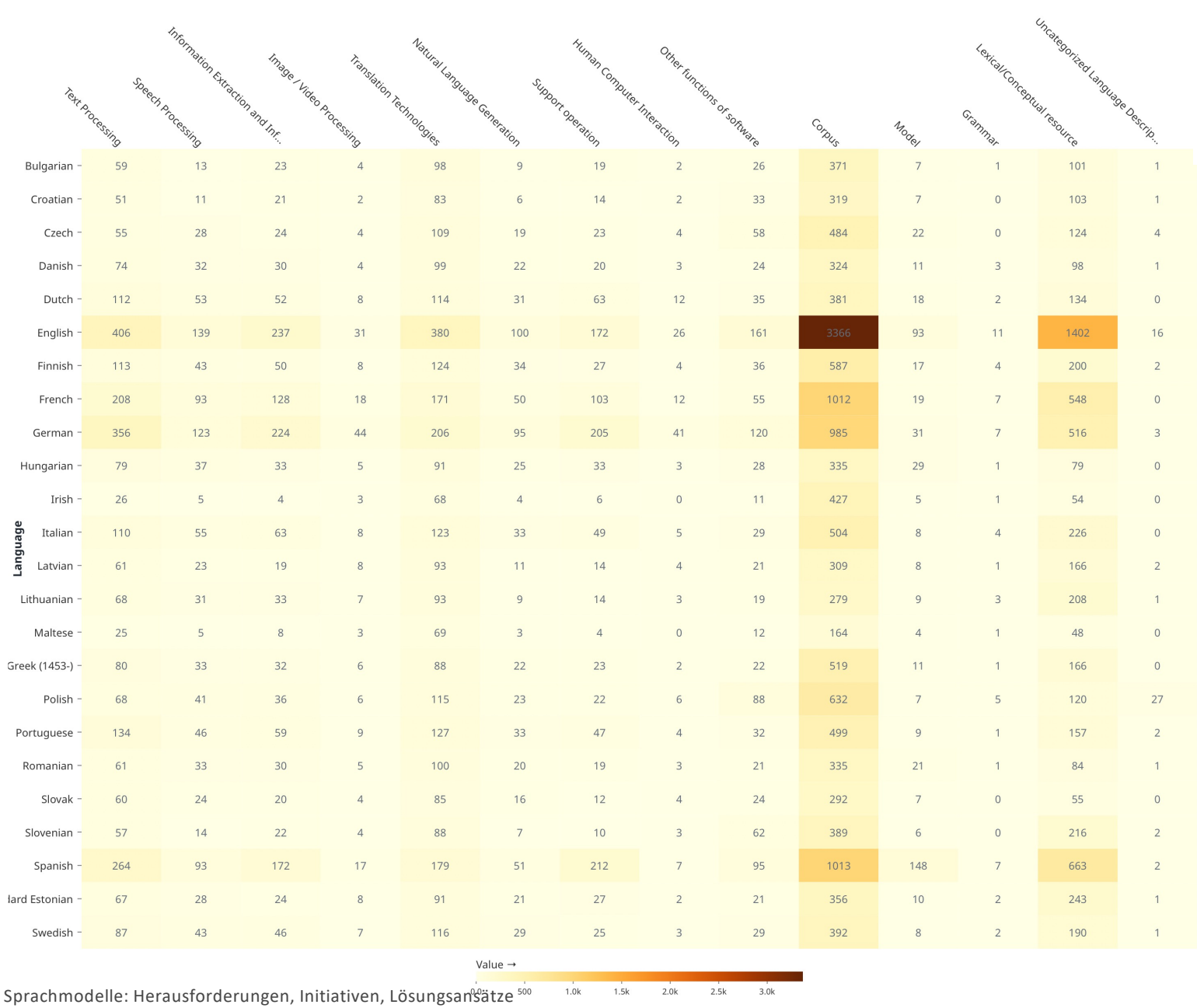
Digital Language Equality Metrik: 2022 vs. 2023 (3/3)

EUROPEAN LANGUAGE EQUALITY

Zeitliche Entwicklung (alternative Visualisierung)

- Alle Sprachen machen Fortschritte, aber …
- Englisch macht deutlich mehr Fortschritte im ersten Halbjahr 2023
- Die Lücke wird größer und nicht etwa kleiner!



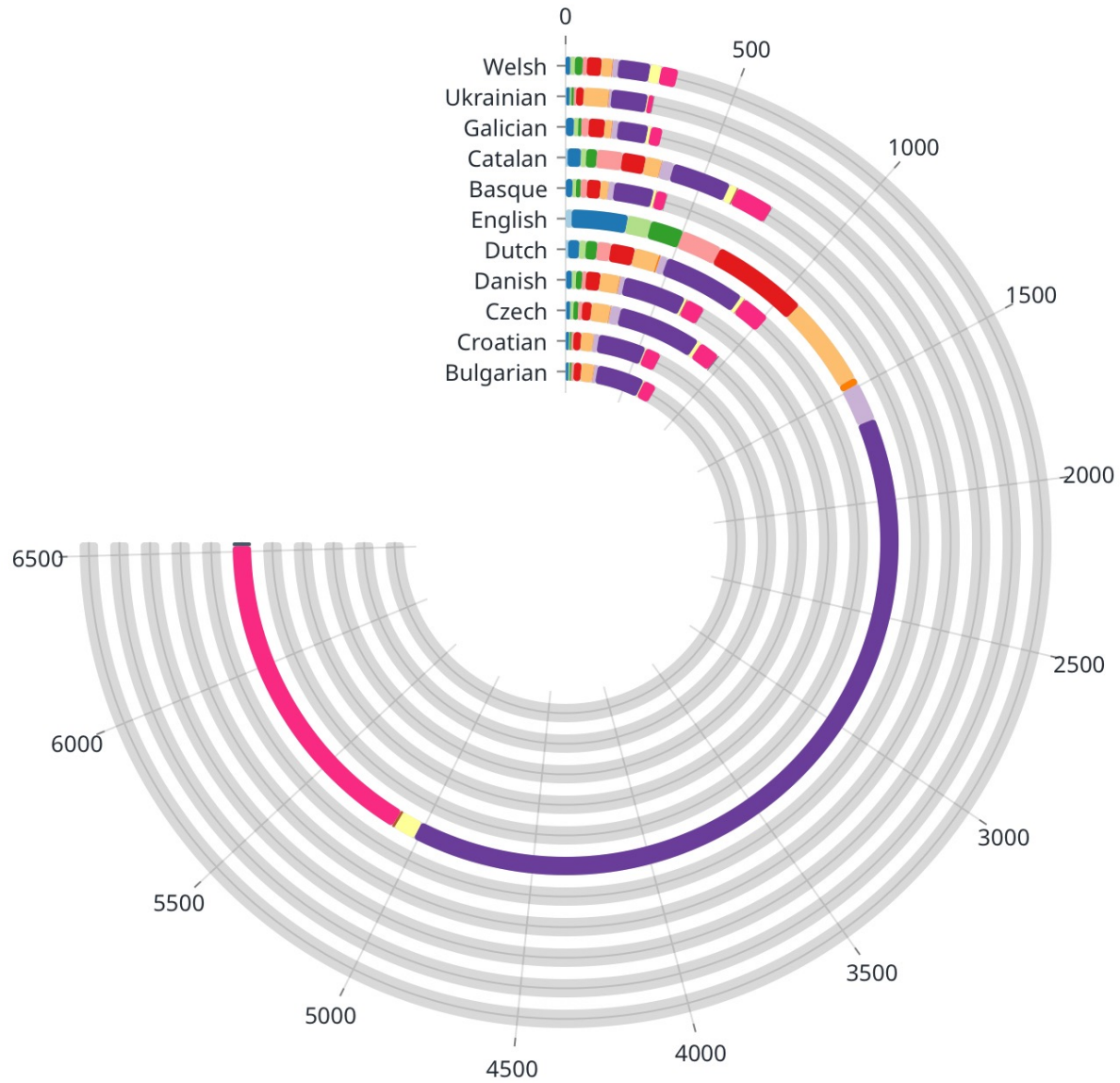


Language	Text Processing	Speech Processing	Information Extraction and Inf...	Image / Video Processing	Translation Technologies	Natural Language Generation	Support operation	Human Computer Interaction	Other functions of software	Corpus	Model	Grammar	Lexical/Conceptual resource	Uncategorized Language Descrip...
Bulgarian	59	13	23	4	98	9	19	2	26	371	7	1	101	1
Croatian	51	11	21	2	83	6	14	2	33	319	7	0	103	1
Czech	55	28	24	4	109	19	23	4	58	484	22	0	124	4
Danish	74	32	30	4	99	22	20	3	24	324	11	3	98	1
Dutch	112	53	52	8	114	31	63	12	35	381	18	2	134	0
English	406	139	237	31	380	100	172	26	161	3366	93	11	1402	16
Finnish	113	43	50	8	124	34	27	4	36	587	17	4	200	2
French	208	93	128	18	171	50	103	12	55	1012	19	7	548	0
German	356	123	224	44	206	95	205	41	120	985	31	7	516	3
Hungarian	79	37	33	5	91	25	33	3	28	335	29	1	79	0
Irish	26	5	4	3	68	4	6	0	11	427	5	1	54	0
Italian	110	55	63	8	123	33	49	5	29	504	8	4	226	0
Latvian	61	23	19	8	93	11	14	4	21	309	8	1	166	2
Lithuanian	68	31	33	7	93	9	14	3	19	279	9	3	208	1
Maltese	25	5	8	3	69	3	4	0	12	164	4	1	48	0
Greek (1453-)	80	33	32	6	88	22	23	2	22	519	11	1	166	0
Polish	68	41	36	6	115	23	22	6	88	632	7	5	120	27
Portuguese	134	46	59	9	127	33	47	4	32	499	9	1	157	2
Romanian	61	33	30	5	100	20	19	3	21	335	21	1	84	1
Slovak	60	24	20	4	85	16	12	4	24	292	7	0	55	0
Slovenian	57	14	22	4	88	7	10	3	62	389	6	0	216	2
Spanish	264	93	172	17	179	51	212	7	95	1013	148	7	663	2
łard Estonian	67	28	24	8	91	21	27	2	21	356	10	2	243	1
Swedish	87	43	46	7	116	29	25	3	29	392	8	2	190	1

EUROPEAN LANGUAGE EQUALITY

Erstellt mit dem ELE-Dashboard

https://live.european-language-grid.eu/catalogue/dashboard



EUROPEAN LANGUAGE EQUALITY

- Human Computer Interaction
- Information Extraction and Information Retrieval
- Natural Language Generation
- Speech Processing
- Support operation
- Text Processing
- Translation Technologies
- Image / Video Processing
- Other functions of software
- Corpus
- Model
- Grammar
- Lexical/Conceptual resource
- Uncategorized Language Description

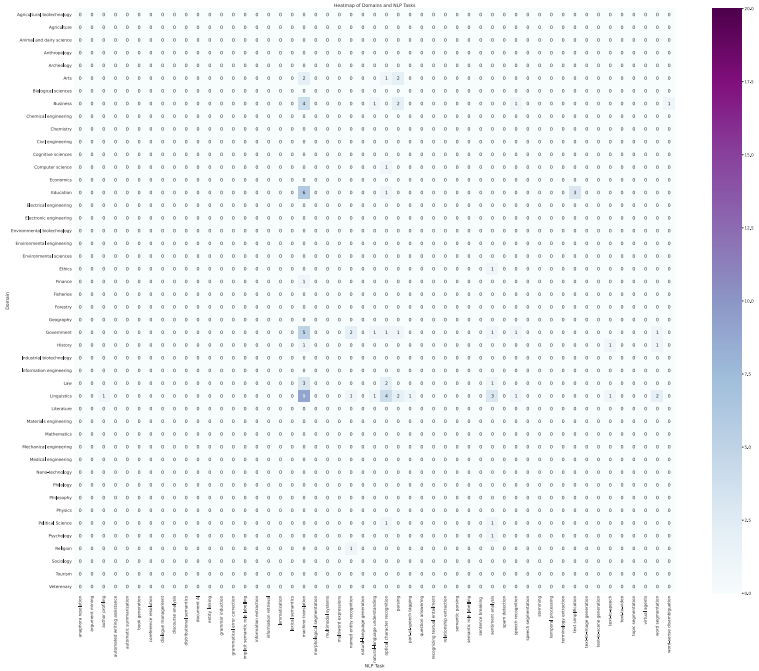
Erstellt mit dem ELE-Dashboard

https://live.european-language-grid.eu/catalogue/dashboard

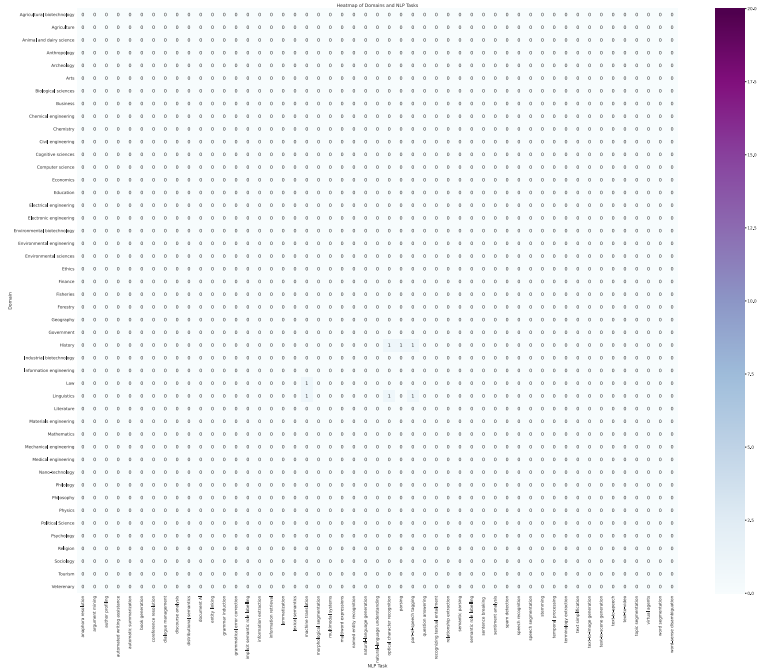
Digital Language *In*equality (Analyse der ACL Anthology)

EUROPEAN LANGUAGE EQUALITY

Englisch



Dänisch



Färöisch

Domains

NLP Tasks

Diego Alves, Marko Tadić: “EuLTDom2023 – European LT Domains 2023” (2023), ELE FSTP Projekt

OpenGPT-X: Große Sprachmodelle für das Deutsche

OpenGPT-X entwickelt große KI-Sprachmodelle, die neue datenbasierte Business-Lösungen ermöglichen und sich speziell an europäische Bedürfnisse richten.

openGPT-X

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

gaia-x

Partner

KI BUNDESVERBAND

ALEPH ALPHA

ControlExpert

DFKI Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Fraunhofer IIS

Fraunhofer IAIS

IONOS

JÜLICH Forschungszentrum

WDR

ZIH Zentrum für Informationsdienste und Hochleistungsrechnen

OpenGPT-X

- Große Sprachmodelle sind mittlerweile eine Schlüsseltechnologie
  - Die erfolgreichsten LLMs stammen aus den USA sowie China.
  - Rechtslage z.B. bzgl. Verarbeitung von Daten, Server-Standorten etc. oftmals unklar.
  - Die rapide zunehmende Relevanz von LLMs bedeutet für Europa:
    - Digital Souveränität: Unabhängigkeit von Technologien und Daten sicherstellen.
    - Mehr Innovation und Wettbewerbsfähigkeit in Europa.
- OpenGPT-X entwickelt offene und europäische Sprachmodelle.
  - Europäischer Datenschutz wird vollständig berücksichtigt.
  - Steigert Innovation und stärkt Europas Wettbewerbsfähigkeit („LLMs made in Europe“).
  - Förderung durch das BMWK (Januar 2022 bis Dezember 2024).

OpenGPT-X: Colossal OSCAR

- Mittlerweile ist klar: Die Trainingsdaten sind entscheidend.
- Eine Hauptquelle sind Webdaten (z.B. Webcrawls von CommonCrawl), da sie – vergleichsweise – unkompliziert riesige Datenmengen verfügbar machen (im Petabyte-Bereich).
- Herausforderung: Was sind gute bzw. schlechte Webdaten?
- OSCAR Projekt (Dr. Pedro Ortiz): Annotierte und gefilterte Webdaten für das Training von Sprachmodellen (z.B. für BLOOM von BigScience)
- Neue Veröffentlichung: Colossal OSCAR (10x mehr Daten – bis zu eine Billionen Tokens inkl. Qualitätsannotationen)
- Mehr als 120TB Daten (150+ Sprachen, unkomprimiert)

OSCAR Projekt
https://oscar-project.org/

OpenGPT-X: lm-datasets

- Neben Webdaten sind kuratierte Datensätze entscheidend für die Fähigkeiten von Sprachmodellen: Qualität statt Masse.
- Kuratierte Datensätze für Sprachmodelle werden bisher nicht einheitlich veröffentlicht (unterschiedliche Formate, Plattformen, Repositorien etc.)
- Das Software-Framework lm-datasets (Dr. Malte Ostendorff) macht viele Datensätze über eine einheitliche Schnittstelle verfügbar.
- Inkl. diverser Werkzeuge für Filterung, Sampling etc.
- Aktuell mehr als 400 Datensätze aus +60 Quellen in 32 Sprachen
- Publikation in Arbeit

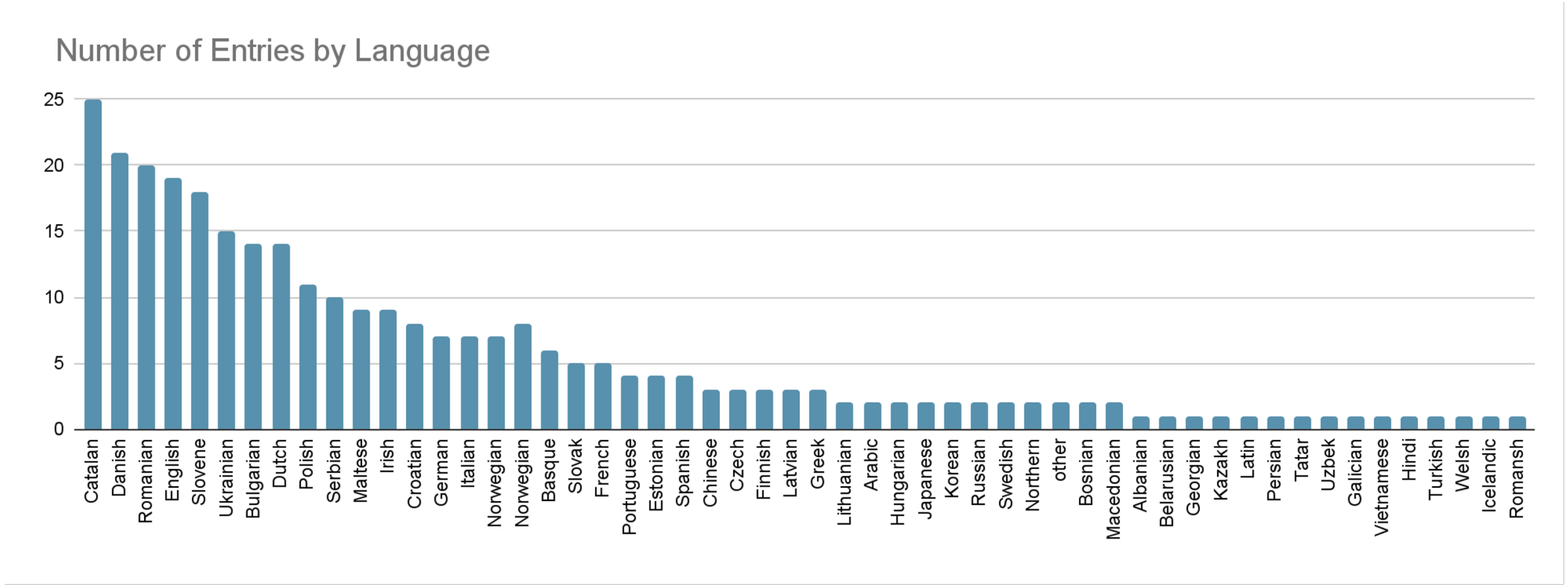
lm-datasets
github.com/malteos/lm-datasets

EUROPEAN LANGUAGE EQUALITY

Zusammenarbeit mit unseren europäischen Partnern

OpenGPT-X: lm-datasets

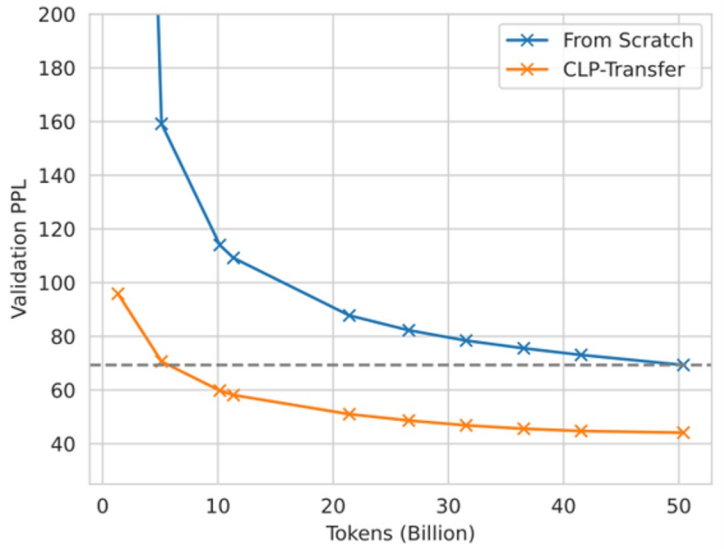
EUROPEAN LANGUAGE EQUALITY

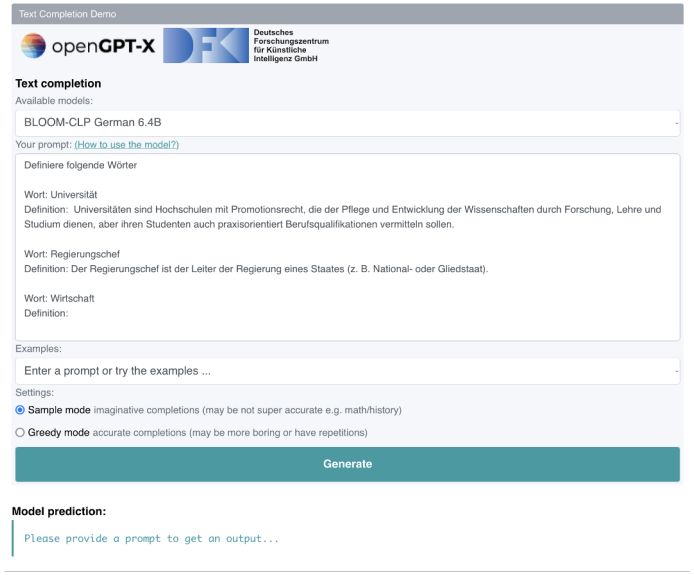


OpenGPT-X: Transfer Learning

- Das “Open” in OpenGPT-X steht auch für die Zugänglichkeit von LLM-Technologie (geringere Barrieren in Bezug auf Ressourcenbedarf)
- Transfer Learning: Bestehende Ressourcen „recylen“ um die Effizienz zu steigern (z.B. Open Source LLMs wie BLOOM anpassen).
- Cross-lingual & Progressive Transfer Learning [1]:
  - Bestehende Sprachmodelle auf eine neue Sprache anpassen (z.B. Englisch → Deutsch)
- Ergebnis: Bis zu 80% Reduktion des Trainingsaufwands.
- Größtes deutsches Open-Source Sprachmodell (bei Publikation):

[BLOOM-CLP-GERMAN (6.4B parameters)]

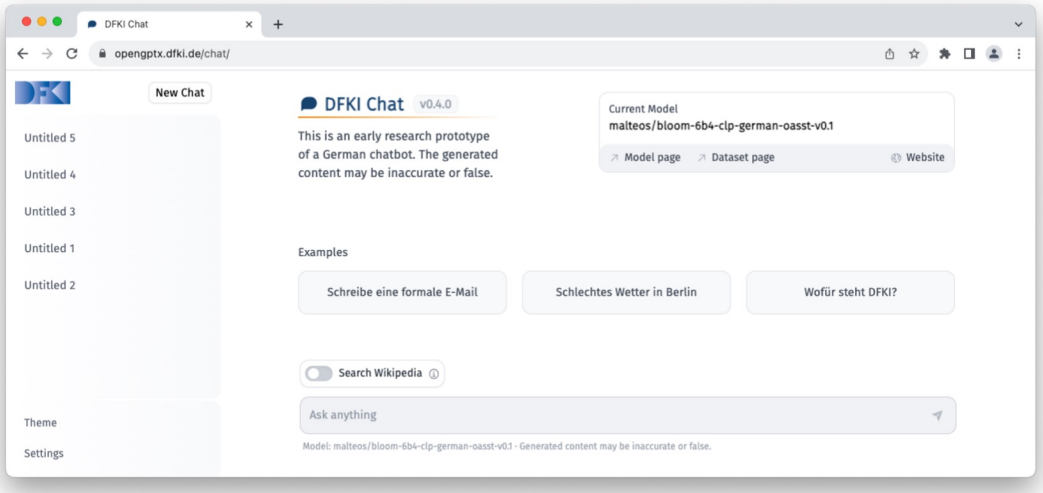




[1] Malte Ostendorff und Georg Rehm. “Efficient Language Model Training through Cross-Lingual and Progressive Transfer Learning”. In: *PML4DC@ICLR 2023.* 05 Mai 2023.

OpenGPT-X: Transfer Learning

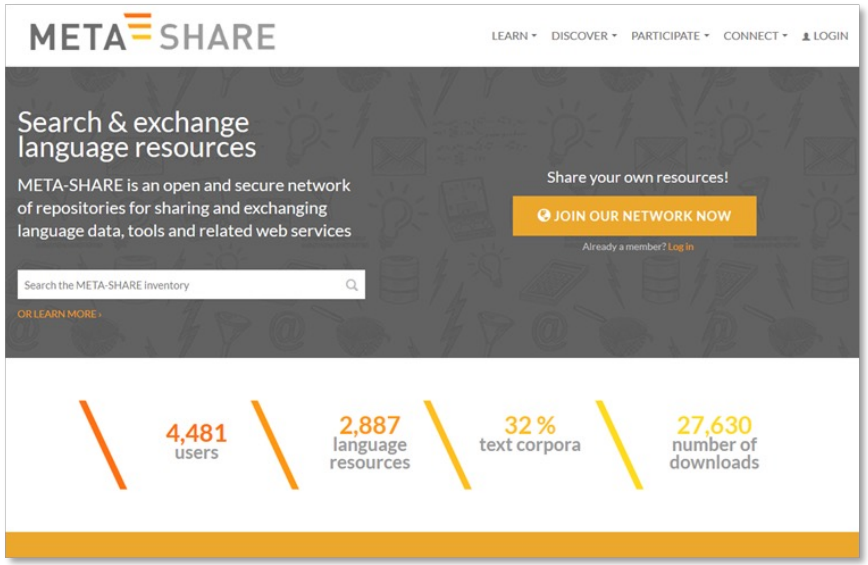
- Zentrale Funktion eines Sprachmodells: Vorhersage des wahrscheinlichsten nächsten Wortes (next word prediction).
- Chatbots oder virtuelle Assistenten (ChatGPT etc.) sind Sprachmodelle, die für Chat optimiert wurden.
- DFKI Chat:
  - Optimierte Variante von BLOOM-CLP-German (siehe vorherige Folie).
  - Integration von Retrieval-Augmented Generation
  - Bei Bedarf können aktuelle Zusatzinformationen aus Wikipedia genutzt werden.

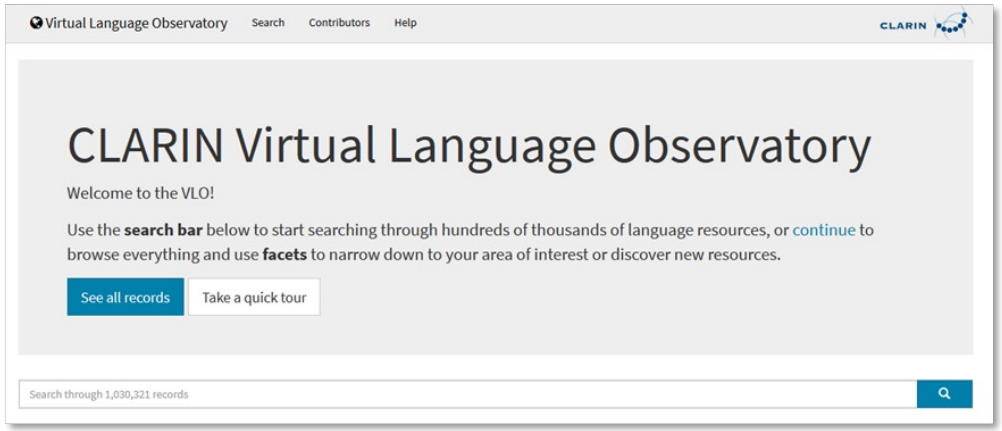


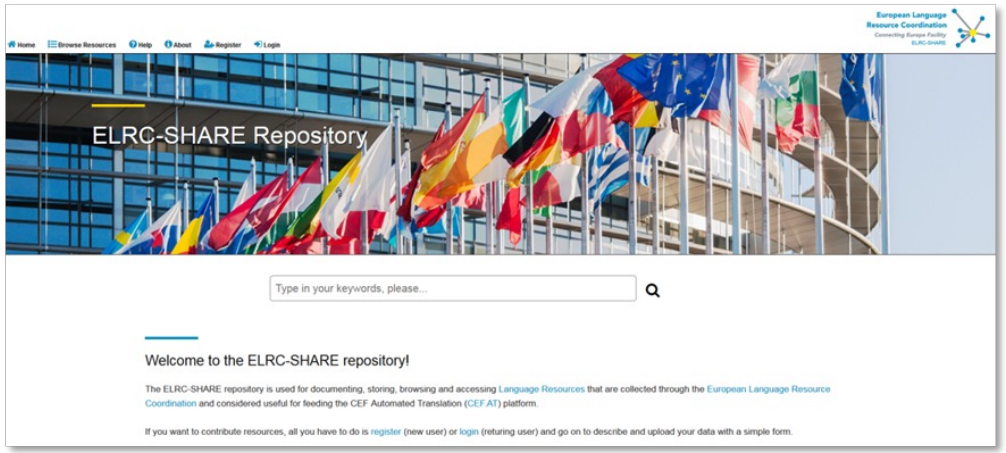
DFKI Chat
https://opengptx.dfki.de/chat

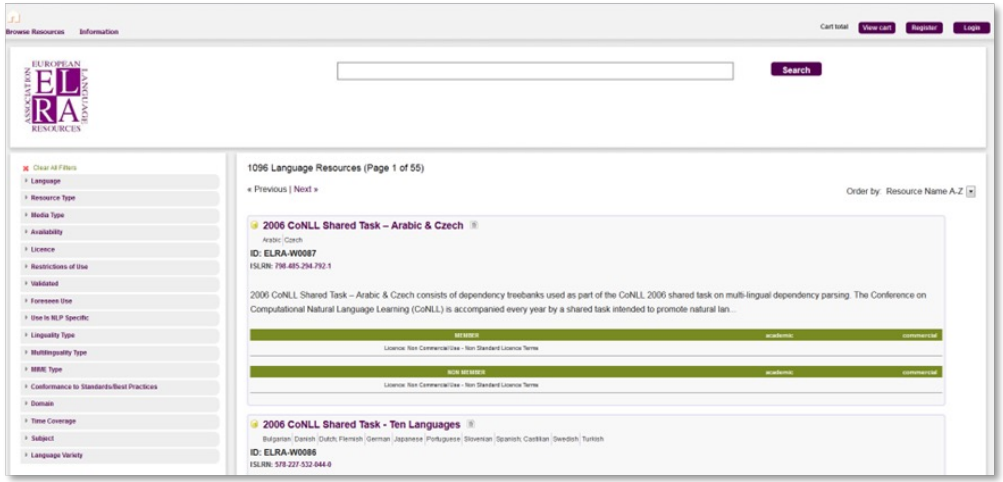
Daten

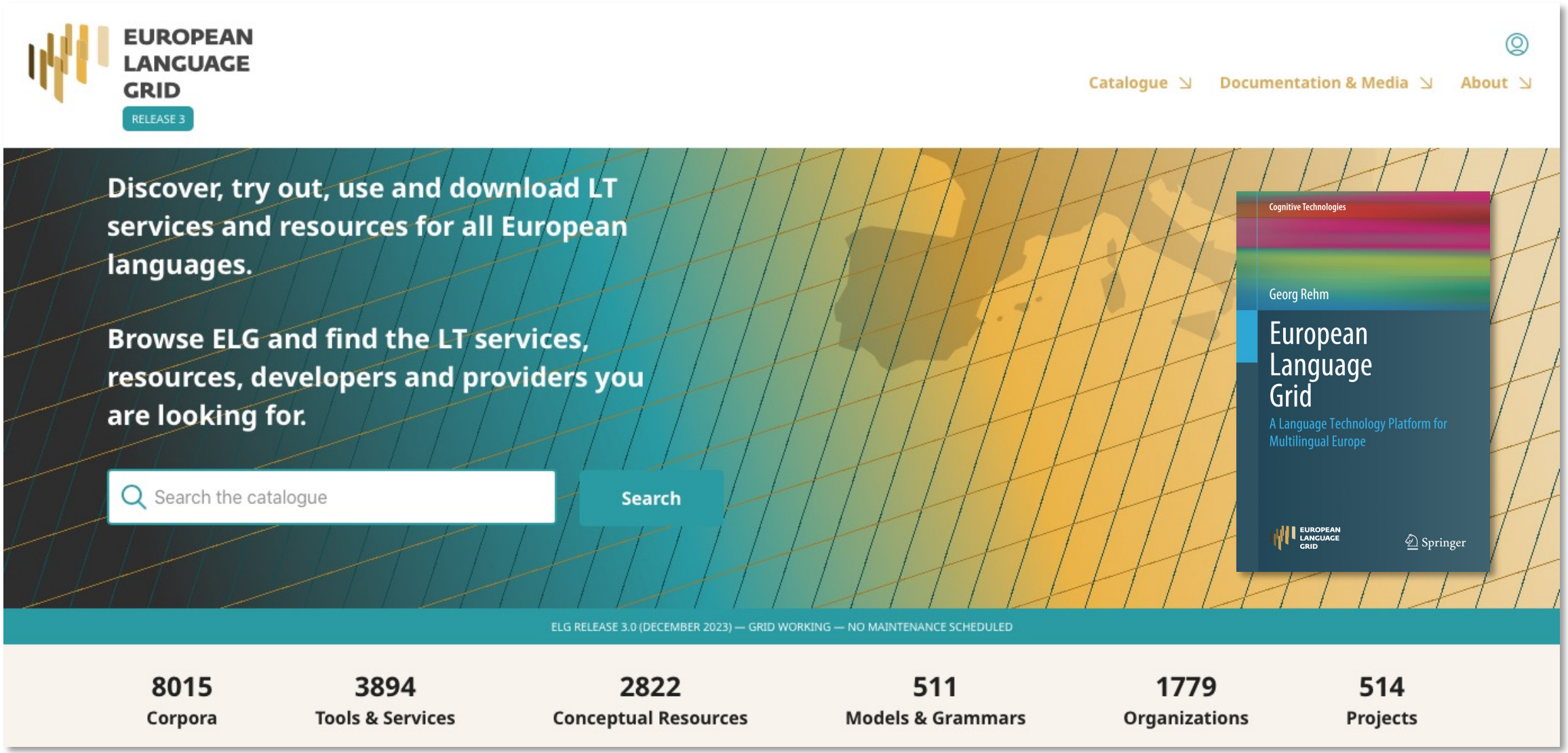
Existierende LT/NLP-Plattformen, -Infrastrukturen, -Repositorien










EUROPEAN LANGUAGE GRID
RELEASE 3
Catalogue
Documentation & Media
About
Discover, try out, use and download LT services and resources for all European languages.
Browse ELG and find the LT services, resources, developers and providers you are looking for.
Search the catalogue
Search
Cognitive Technologies
Georg Rehm
European Language Grid
A Language Technology Platform for Multilingual Europe
EUROPEAN LANGUAGE GRID
Springer
ELG RELEASE 3.0 (DECEMBER 2023) — GRID WORKING — NO MAINTENANCE SCHEDULED
8015 Corpora
3894 Tools & Services
2822 Conceptual Resources
511 Models & Grammars
1779 Organizations
514 Projects

EU Datenstrategie und Datenräume (Data Spaces)

- Data Spaces sind inhärenter Teil der EU-Datenstrategie
- Data Spaces sind auch ein Instrument, um eine Datenwirtschaft in Europa zu etablieren
- Diese Daten können für diverse Zwecke verwendet werden, u.a. auch für das Training von LLMs
- In Europa existieren verschiedene Initiativen bzgl. Datenwirtschaft und Dateninfrastrukturen, die ähnliche Ziele und Positionen haben, sich aber teilweise unterscheiden:
  - Data Spaces Business Alliance (DSBA): Gaia-X, IDSA, FIWARE, BDVA – *Konvergenz!*
  - EU: DSSC (incl. DSBA), SIMPL, 14 offizielle EU Data Space-Projekte
- Der Common European Language Data Space ist einer dieser 14 offiziellen EU Data Spaces.

Common European Language Data Space

- Procurement Projekt (CNECT/LUX/2022/OP/0026)
- Budget: 6M€ (+ 2M€, sofern die EU die Option der Verlängerung wählt)
- Laufzeit: 36 Monate (+ 12 Monate)
- Ziel: Entwicklung und Bereitstellung einer europäischen Plattform mit Marktplatz für die Sammlung, Erstellung, Teilen und Nutzung mehrsprachiger und multimodaler Sprachdaten
- Wichtige Komponenten: Governance, Architektur, Infrastruktur, Offenheit, Dissemination
- Stakeholder: Industrie, Forschung, öffentliche Verwaltung, Kultureinrichtungen, NGOs etc.

Konsortium und Unterauftragnehmer

Koordinator		
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH	DFKI	DE
Partner		
R.C. “Athena”, Institute for Language and Speech Processing	ILSP	GR
Evaluations and Language Resources Distribution Agency	ELDA	FR
TILDE	TILDE	LV
Unterauftragnehmer		
3pc GmbH Neue Kommunikation	3pc	DE
Capgemini Deutschland GmbH	CapG	DE
CLARIN ERIC	CLARIN	NL
Big Data Value Association (Data, AI and Robotics) AISBL	BDVA	BE

Außerdem: Rechtsexperten (Delcade, Frankreich) und ca. 30 Organisationen, die die vorgesehenen Workshops in den Ländern organisieren.

Frühere Projekte und Initiativen

- Die Kerngruppe – DFKI, ILSP, ELDA, TILDE – arbeitet seit Jahren in diversen Projekten bzgl. Infrastruktur und Plattformen zusammen:
- **META-NET** (FP7, 2010-2013)
  - META-SHARE
- **ELRC** (CEF, 2014-2023)
  - ELRC-SHARE
- **ELG** (H2020, 2019-2022)
  - ELG Cloud Platform
- **ELE** (PP/PA, 2021-2023)

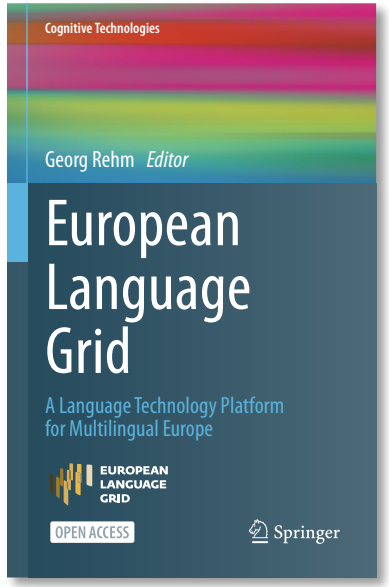
META-NET

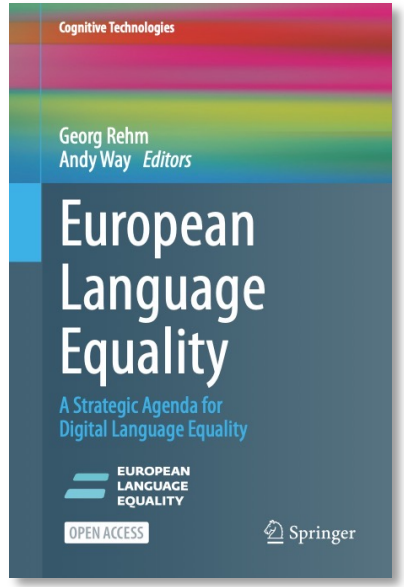
European Language Resource Coordination
Connecting Europe Facility

EUROPEAN LANGUAGE GRID

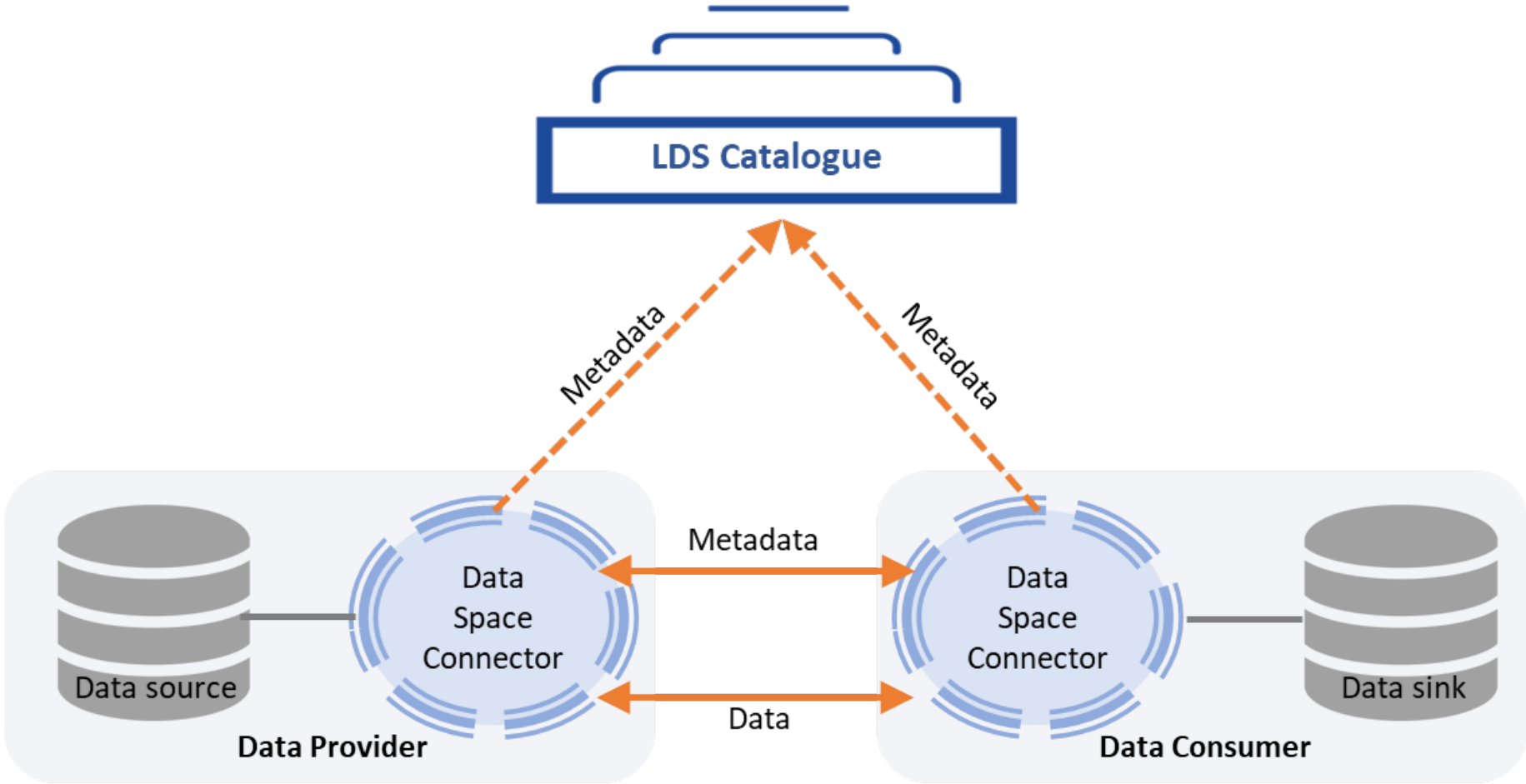
EUROPEAN LANGUAGE EQUALITY

Die **dezentrale LDS-Infrastruktur** wird auf Erfahrungen aus ELG, ELRC-SHARE und META-SHARE basieren.

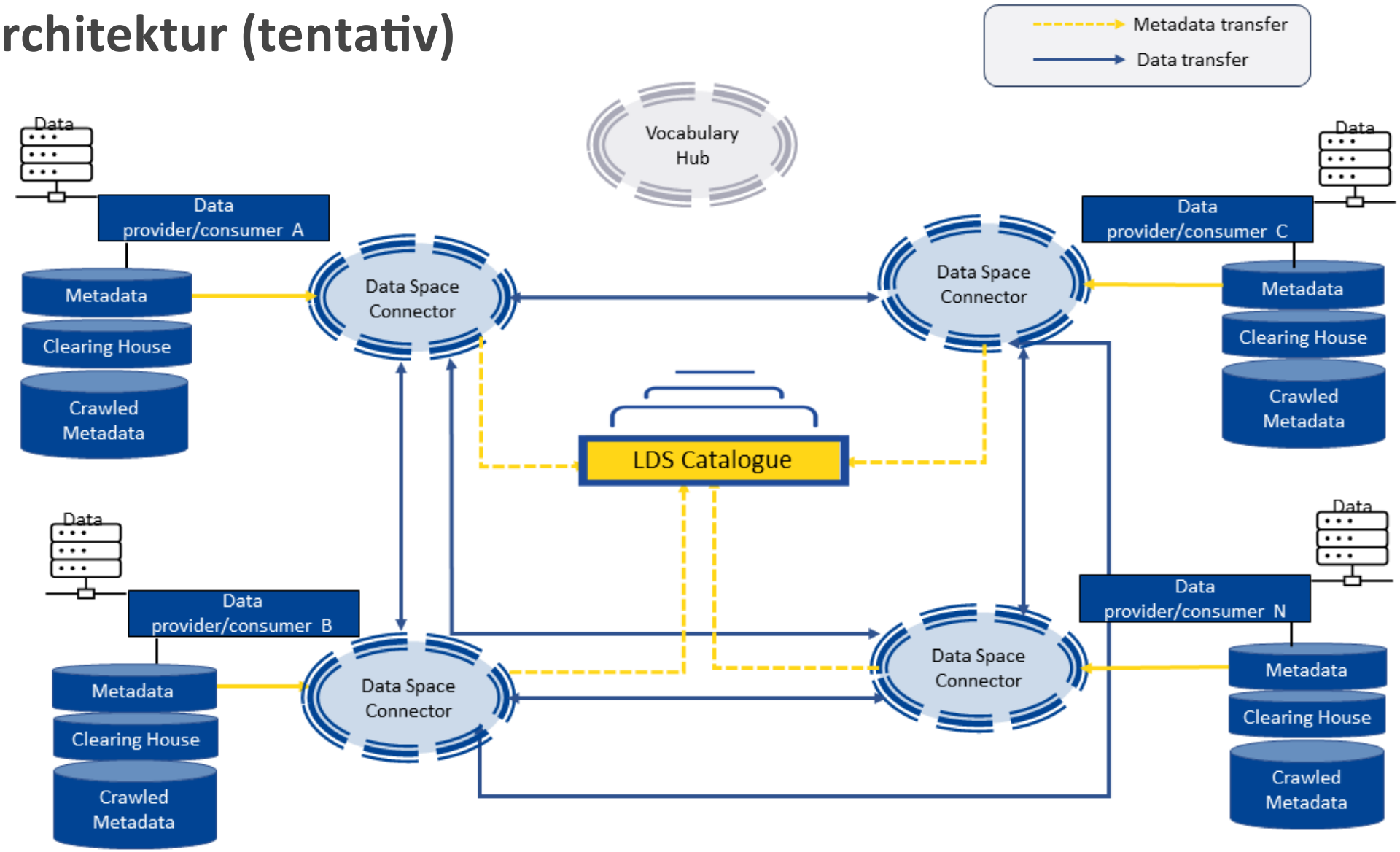




LDS Datenflüsse



LDS Architektur (tentativ)



Unterschiedliche Klassen von Daten

Klasse	Typische Größe	Anbieter	Integration in LDS	Relevanz für LLMs
Reguläre Korpora & Sprachressourcen	Klein (MB, GB)	Primär NLP/LT-Forschung: ELG, META-SHARE, CLARIN, ELRA, ELDA etc.	Können einfach integriert werden, indem die Repositorien an LDS angebunden werden	Typischerweise qualitativ sehr hochwertige Daten und daher relevant für LLMs, aber nicht als “Basisdaten” für das Training
Web-Crawls	Sehr groß (TB, PB)	Common Crawl (und OSCAR-Dumps), Internet Archive-Dumps etc.	Herausforderung ist die Größe (aufwendig zu kopieren, zu prozessieren, zu speichern; müssen sehr nah an HPC sein)	Unabdingbar auf Grund ihrer Größe und Abdeckung – aber: viel Rauschen, großer Bedarf für Vorverarbeitung und Filterung
Neue, frische Daten aus der Industrie sowie von anderen Organisationen	Beliebige Größe, idealerweise so groß wie möglich	Verlage, Medienanbieter, Bibliotheken, Call-Centres, Radio- und Fernsehanstal-ten, andere Data Spaces	Können einfach integriert werden, indem die Organisationen an LDS angebunden werden	Insbesondere hochqualitative Daten oder domänenspezifische Daten oder Daten, die bestimmte Sprachen abdecken, daher sehr relevant für LLMs

Wissenschaftliche Daten

NFDI for Data Science and Artificial Intelligence

- NFDI: Nationale Forschungsdateninfrastruktur
- NFDI4DS: NFDI for Data Science and Artificial Intelligence (Laufzeit 5 + 5 Jahre)
- NFDI4DS entwickelt eine Forschungsdateninfrastruktur für DS und KI
- Basis: Existierende Infrastrukturen und Werkzeuge, die interoperabel gemacht und anschließend verbunden werden (inklusive u.a. European Language Grid)
- Langfristig wird die NFDI4DS-Infrastruktur mit der restlichen NFDI-Infrastruktur verbunden
- DFKI co-koordiniert zwei der fünf Task Areas:
  - Infrastructure and Services
  - Data Science and AI – Transfer and Application
- Forschungsthemen: Scholarly Information Processing, Informationsextraktion, Erstellung von Research Knowledge Graphs (RKGs)

NFDI for Data Science and Artificial Intelligence

- Es existieren diverse Initiativen zur Repräsentation wissenschaftlichen Wissens in Form von Knowledge Graphs (Research Knowledge Graphs, RKGs), z.B. ORKG (TIB Hannover)
- In NFDI4DS konzentrieren wir uns auf die Verarbeitung wissenschaftlicher Publikationen
- Domäne Computerlinguistik, d.h. Publikationen der Association for Computational Linguistics (ACL)
- Aktuell zwei Ziele:
  - Klassifikation beliebiger wissenschaftlicher Artikel in ihre Fachrichtungen
  - Extraktion von Informationen aus Publikationen und Integration dieser Informationen in RKGs
- Vision: Robuste Extraktion derartiger Informationen aus beliebigen wissenschaftlichen Artikeln und ihre Integration in wissenschaftliche Wissensgraphen
- *RKGs repräsentieren wissenschaftliche Ergebnisse, Daten und Artefakte explizit (symbolisch)*
- *LLMs repräsentieren wissenschaftliche Ergebnisse, Daten und Artefakte implizit (subsymbolisch)*
- Ein hybrider Ansatz erscheint sinnvoll, d.h. *LLMs für RKGs* und *RKGs für LLMs*

Schlussfolgerungen

Zusammenfassung

- LLMs sind nicht für Europas Sprachen und ihre linguistischen und kulturelle Bedürfnisse optimiert.
- Das DFKI arbeitet in diversen Projekten an europäischen Sprachdaten und Sprachmodellen.
- Erste Modelle wurden bereits veröffentlicht – weitere folgen Anfang 2024.
- Der Language Data Space läuft auf Hochtouren: Technologie, Dissemination, Governance etc.
- Zusammenarbeit mit allen relevanten Initiativen: DSSC, HPLT, EuroHPC, OpenWebSearch etc.
- Von zentraler Bedeutung: Nutzung des LDS durch Industrie und andere Organisationen, die in sinnvoller Weise Daten bereitstellen und teilen können sowie auch wollen
- Ziel: Identifizierung und Bereitstellung neuer und „frischer“ Sprachdaten, speziell aus der Industrie sowie von anderen Organisationen, die interessante Daten haben, und idealerweise auch viele bzw. alle europäischen Sprachen sowie auch alle Modalitäten abdecken

Zusammenfassung

- Modelle:
  - Diverse LLM-Forschungprojekte in Europa – jeweils mit zwei Jahren administrativem Vorlauf
  - Diverse LLM-Firmen in Europa versuchen, mit OpenAI zu konkurrieren – Herausforderung!
- Daten:
  - In Europa ist digitale Sprachgerechtigkeit noch in weiter Ferne
  - Common European Language Data Space
  - Erste nationale Language Data Spaces entstehen, z.B. in Spanien
- Wissenschaftliche Daten:
  - Diverse RKGs/SKGs werden aktiv mit Inhalten gefüllt
- *Erstmals institutionelle europäische Kollaboration!* (ab Anfang 2024):
  - ALT-EDIC

Alliance for Language Technologies EDIC (ALT-EDIC)

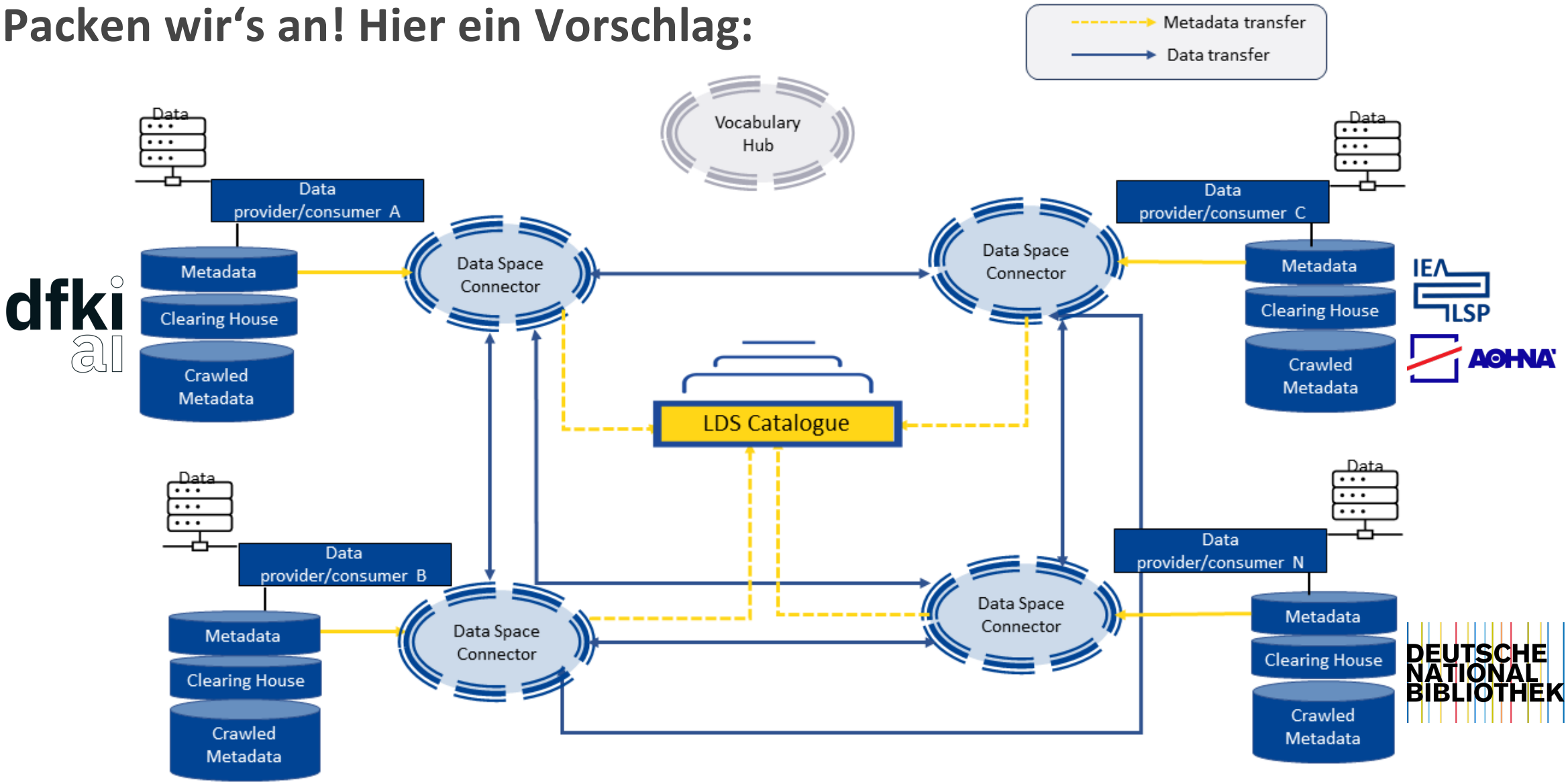
- European Digital Infrastructure Consortium (EDIC): Ein neuer Typ juristische Person in der EU
- 15+ EU-Mitgliedsstaaten sind aktuell an diesem Multi-Country Project (MCP) beteiligt
- Koordiniert vom französischen Ministerium für Kultur
- Enge Zusammenarbeit zwischen: ALT-EDIC Arbeitsgruppe, EC, LDS
- ALT-EDIC Aktionsplan konzentriert sich auf:
  1. Daten;
  2. Existierende und neue Sprachmodelle;
  3. Evaluatiering, Zertifizierung, Normalisierung;
  4. Ökosystem;
  5. EDIC-Implementierung

KI in Bibliotheken – Eine Vision – Schlussfolgerungen

**Mittels ScienceGPT mit der wissenschaftlichen Literatur sprechen,
mit allen wissenschaftlichen Publikationen in den Dialog treten,
diese aktiv konsultieren, Forschung gemeinsam mit den Quellen betreiben!**

- In dieser Vision liefern Bibliotheken nicht mehr „nur“ Hinweise auf relevante Literatur (und die Literatur selbst), sondern Nutzerinnen können sich aktiv mit der wissenschaftlichen Literatur „unterhalten“, extrem effizient recherchieren und kreativ inspirieren lassen.
- Ist das realistisch? Ja!
- Werden sich Bibliotheken grundsätzlich transformieren ... müssen? Vermutlich!
- Durch systematische Datenbereitstellung und Datennutzung können Bibliotheken und wissenschaftliche Literatur eine Schlüsselrolle für die nächsten Durchbrüche bei LLMs spielen.

Packen wir‘s an! Hier ein Vorschlag:



Herzlichen Dank!

A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

Prof. Dr. Georg Rehm (DFKI GmbH)
georg.rehm@dfki.de

07.12.2023 – KI in Bibliotheken: Neue Wege mit großen Sprachmodellen? – DNB, Frankfurt
https://language-data-space.ec.europa.eu