Erschienen in: Dąbrowska-Burkhardt, Jarochna/Eichinger, Ludwig M./Itakura, Uta (Hrsg.): Deutsch: lokal – regional – global. Tübingen: Narr, 2017. S. 103-113. (Studien zur Deutschen Sprache 77)

MONICA FÜRBACHER / TAMÁS VÁRADI / ANDREAS WITT

DIGITALE FORSCHUNGSINFRASTRUKTUREN: IHRE NUTZUNG DURCH DIE MITGLIEDER DER EUROPÄISCHEN FÖDERATION NATIONALER SPRACHINSTITUTIONEN

1. Einleitung

Professor Stickel war von 1976 bis 2002 Direktor des Instituts für Deutsche Sprache, einer zentralen außeruniversitären Institution, deren Aufgabe die Erforschung und Dokumentation der deutschen Sprache in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte ist.

Nach seinem Ausscheiden widmete er sich dem Aufbau einer Föderation, die sich mit der Bewahrung der europäischen Sprachen und der damit einhergehenden sprachlichen Vielfalt beschäftigt.

Ein herausragendes Ergebnis dieser Aktivitäten war die Konstituierung der unabhängigen Organisation *European Federation of National Institutions for Language* (EFNIL),[1] die am 14. Oktober 2003 von 12 Vertretern aus Institutionen der EU-Mitgliedsstaaten in Stockholm gegründet wurde. EFNIL vereint zentrale nationale wie internationale Institutionen innerhalb der Europäischen Union (EU), die sich zur Aufgabe gemacht haben, die Landessprache(n) nicht nur zu beobachten und zu analysieren, sondern sprachliche Ressourcen zur Verfügung zu stellen oder gar sprachpolitische Richtlinien zu konzipieren, um den Erhalt der jeweiligen Sprache zu gewährleisten. Zurzeit besteht die Föderation aus 27 Mitgliedern sowie aus zwei assoziierten Mitgliedern.

EFNIL bietet diesen Instituten eine Plattform, zum einen um Informationen über die jeweiligen, offiziell anerkannten (Standard-)Sprachen auszutauschen, zum anderen aber auch zur gegenseitigen Unterstützung von sprachpolitischen Richtlinien in der EU. Daneben hat sich EFNIL zum Ziel gesetzt, die linguistische sowie kulturelle Diversität innerhalb der EU zu bewahren und zu fördern. Der Informationsaustausch zwischen den EFNIL-Mitgliedern und die internationale Sichtbarkeit dieser Föderation werden u.a. durch die jährlich stattfindende Tagung gefördert, die Verabschiedung von Resolutionen und die Publikation von Jahrbüchern, die über die EFNIL-Webseite frei zugänglich sind.

---

[1] www.efnil.org (Stand: 18.11.2016).

Eines der EFNIL-Mitglieder ist das Institut für Deutsche Sprache, nicht zuletzt dank Gerhard Stickel, der als einer der Gründungsmitglieder von der ersten Stunde an mit der Föderation vertraut ist. Bis heute ist Professor Stickel Präsident der EFNIL und vertritt als Repräsentant das IDS.

Innerhalb EFNILs wird die linguistische Situation in der EU beschrieben und wissenschaftlich analysiert und es werden darüber hinaus cross-linguale Probleme sowie Schwierigkeiten hinsichtlich der sprachpolitischen Standardisierung eingehend geprüft. Des Weiteren unterstützt EFNIL die Untersuchung der offiziellen europäischen Sprachen, um die schriftliche und mündliche Qualifikation zu fördern, indem sie Nicht-Muttersprachlern ermöglicht, die jeweilige(n) (National-)Sprache(n) zu lernen und sich mit Studierenden und Lehrenden innerhalb der EU auszutauschen.

Die zunehmende Bedeutung der Forschungsinfrastruktur offenbarte sich bereits bei der Jahrestagung 2010 „Language, Languages and New Technologies", bei der sich die EFNIL-Mitglieder mit dem Einsatz und der Nutzung von Informationstechnologien für eine verbesserte Sprachforschung und -dokumentation innerhalb der europäischen Länder auseinandersetzen und ihre Entwicklung, ihre aktuelle Situation und ihre Zukunftsperspektive analysierten.

2. (Weiter-)Entwicklung von (elektronischen) Sprachdaten

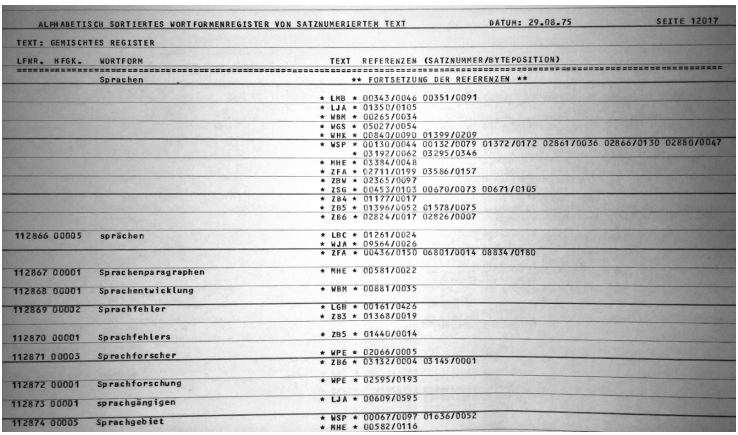
Schon früh wurden Daten erhoben und gesammelt, um die Sprache zu analysieren, den Sprachwandel nachvollziehen zu können sowie Muster und Regeln in ihr zu erkennen, aber auch zur Belegsammlung. Bereits 1854 veröffentlichten die Gebrüder Grimm das erste *Deutsche Wörterbuch* (DWB), das von einer akribischen Belegsammlung der deutschen Sprache zeugt.[2] Auch in anderen Ländern wurden textuelle Daten als Belege gesammelt, so wurde der erste Teil des *Oxford English Dictionary* (OED) bereits 1884 veröffentlicht – der vierte Teil 1928.[3]

Während Texte vormals gedruckt vorlagen und manuell gesichtet wurden, begann man schließlich diese elektronisch zu sammeln. So wurde bereits Anfang der 1960er Jahre an der Brown University das erste, heutzutage als *Brown Corpus* bekannte Korpus von W. Nelson Francis und Henry Kučera veröffentlicht. Ihr Korpus des amerikanischen Englisch der Gegenwartssprache umfasste 500 Belegtexte mit ca. 2.000 Wörtern und mit bereits damals insgesamt ca. 1 Million Wörtern (Francis/Kučera 1979).

---

[2] Mittlerweile ist das DWB auch online zugänglich: http://woerterbuchnetz.de/DWB (Stand: 18.11.2016).

[3] Im OED lässt sich auch online recherchieren: www.oed.com (Stand: 18.11.2016).



ALPHABETISCH SORTIERTES WORTFORMENREGISTER VON SATZNUMERIERTEM TEXT DATUM: 29.08.75 SEITE 12017

TEXT: GEMISCHTES REGISTER

LFNR.	HFGK.	WORTFORM	TEXT REFERENZEN (SATZNUMMER/BYTEPOSITION)
		Sprachen	** FORTSETZUNG DER REFERENZEN **
			* LMB * 00343/0046 00351/0091
			* LJA * 01350/0105
			* WBK * 00265/0034
			* WGS * 05027/0054
			* WMK * 00840/0090 01399/0209
			* WSP * 00130/0044 00132/0079 013727/0172 028617/0036 028667/0130 02880/0047
			* 03192/0062 03295/0346
			* MHE * 03384/0048
			* ZFA * 02711/0199 03586/0157
			* ZBN * 02365/0097
			* ZSG * 00453/0103 00670/0073 00671/0105
			* ZB4 * 01177/0017
			* ZB5 * 01396/0052 01578/0075
			* ZB6 * 02824/0017 02826/0007
112866	00005	sprächen	* LBC * 01261/0024
			* WJA * 09564/0026
			* ZFA * 00436/0150 06801/0014 08834/0180
112867	00001	Sprachenparagraphen	* MHE * 00581/0022
112868	00001	Sprachentwicklung	* WBK * 00881/0035
112869	00002	Sprachfehler	* LGB * 00161/0426
			* ZB3 * 01368/0019
112870	00001	Sprachfehlers	* ZB5 * 01440/0014
112871	00003	Sprachforscher	* WPE * 02066/0005
			* ZB6 * 03132/0004 03145/0001
112872	00001	Sprachforschung	* WPE * 02595/0193
112873	00001	sprachgängigen	* LJA * 06069/0595
112874	00005	Sprachgebiet	* WSP * 00067/0097 01636/0052
			* MHE * 00582/0116

Abb. 1: Auszug eines Wortformenregisters aus dem Jahr 1975

Mit der elektronischen Erfassung von Sprachdaten begann man kurz darauf auch in Deutschland. So fing das Institut für Deutsche Sprache (IDS) kurz nach seiner Gründung in den 1960er Jahren an, deutschsprachige Texte elektronisch auf Lochkarten zu sammeln. Mit einem Umfang von 293 Texten, die zusammen 2,2 Millionen Wörter beinhalteten, veröffentlichte das IDS 1969 mit dem *Mannheimer Korpus I* (MK I) das erste elektronische Korpus deutschsprachiger Texte. Es bestand zum damaligen Zeitpunkt hauptsächlich aus literarischen, aber auch aus populärwissenschaftlichen Werken sowie Zeitungstexten (Teubert/Belica 2014). Im Jahr 1975 wurden einige der IDS-Korpora auf Endlospapier gedruckt und im IDS archiviert. Auch das *LIMAS-Korpus* (Linguistik und Maschinelle Sprachbearbeitung), in dem von 1970 bis 1971 zu 33 Themenbereichen 500 Textstücke mit je 2.000 Textwörtern festgehalten wurden, befindet sich bis heute im Archiv des IDS.

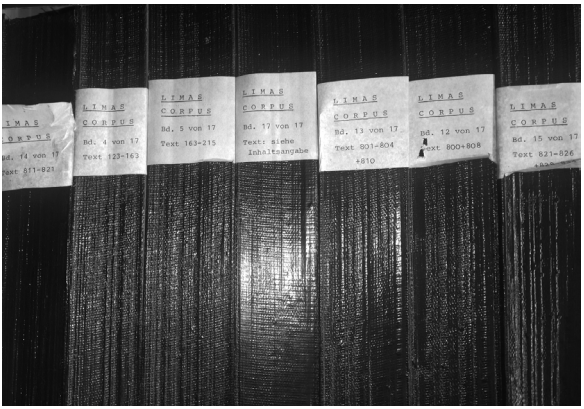


Abb. 2: Ein Teil des LIMAS-Korpus

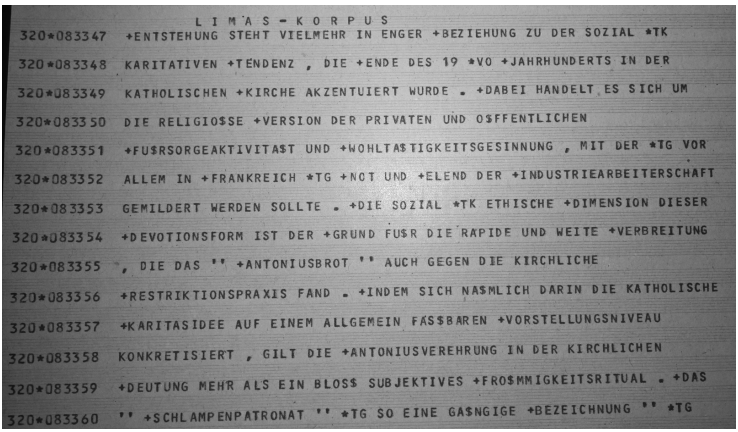


Abb. 3: Belegtext aus dem LIMAS-Korpus

Neben der Sammlung schriftlicher Sprachdaten wurden auch Aufnahmen von gesprochener Sprache sowie deren Transkriptionen gesammelt. Bereits 1932 wurde das *Deutsche Spracharchiv* (DSAv) gegründet, welches 1971 in das IDS integriert wurde. Das DSAv wurde 2004 in *Archiv für Gesprochenes Deutsch* (AGD) umbenannt und wird bis heute weiter ausgebaut.

In den darauffolgenden Jahrzehnten kamen weitere schrift- und gesprochensprachliche sowie multimodale Korpora hinzu, die für unterschiedliche Zwecke entwickelt wurden. Dabei existieren neben den klassischen Referenzkorpora auch Lernerkorpora, Parallel- und Vergleichskorpora sowie genrespezifische und historische Korpora, um nur ein paar zu nennen. Die Korpora unterscheiden sich dabei nicht nur von ihrer Art und Umfang, sondern sie sind verschiedenartig getaggt. Dadurch werden Korpora auf immer mehr Gebieten zur Thesenstützung oder -überprüfung verwendet, wie beispielsweise in der theoretischen/quantitativen Linguistik, in der Grammatikographie wie auch in der Lexikographie, der Translationswissenschaft sowie in der Computerlinguistik.

Sinn und Zweck solch großer Korpora ist es, eine zuverlässigere Beschreibung der deutschen Standardsprache, Dialekte und regionale Unterschiede sowohl in der Vergangenheit als auch in der Gegenwart zu erhalten, um so germanistisch-linguistische Fragestellungen zu lösen.

3. Digitale Sprachressourcen heute

Mittlerweile existieren verschiedene Arten von digitalen Sprachressourcen. Abgesehen von großen Korpora, bestehend aus verschiedenen Archiven und Unterarchiven, wurden auch spezialisierte Korpora, z.B. genre- oder regio-

nalspezifische, entwickelt. Neben maschinell getaggten existieren auch manuell annotierte Korpora, die jeweils unterschiedliche (Such-)Funktionen ermöglichen. Darüber hinaus gibt es eine Vielzahl an lexikalisch-semantischen Netzen, wie beispielsweise GermaNet,[4] WordNet[5] oder TemaNet.[6]

Der Zugang zu diesen digitalen Sprachressourcen erfolgt unterschiedlich. Ein Teil der Ressourcen ist online frei zugänglich oder erfordert eine Registrierung, nach der man unter Umständen dennoch einen eingeschränkten Zugriff auf die unterschiedlichen Archive besitzt, während ein anderer Teil kostenpflichtig ist. Grund dafür sind hauptsächlich rechtliche Belange. Darüber hinaus sind viele der Ressourcen weit verteilt und nur schwer auffindbar.

Mit mittlerweile über 29 Milliarden Wörtern zählt das am IDS verankerte *Deutsche Referenzkorpus* (DeReKo) als die weltweit größte Datensammlung deutschsprachiger Texte. Das DeReKo gliedert sich in mehrere Korpora, auf die sich nach einmaliger, kostenloser Registrierung über die Schnittstelle *COSMAS II*[7] (Corpus Search, Management and Analysis System) zugreifen lässt. Die komplette Datensammlung ist dabei nicht nur mit Metadaten, wie beispielsweise Autor, Verlag und Entstehungsort, versehen, sondern auch mit dem TreeTagger,[8] dem Machinese Phrase Tagger[9] und anderen Werkzeugen morphosyntaktisch annotiert, wodurch Suchanfragen unterschiedlicher Granularität zur Findung grammatischer Muster ermöglicht wird.

Das an der Ungarischen Akademie entwickelte *Hungarian National Corpus* (HNC)[10] verfügt derzeit über ca. 187 Millionen Wörter (Stand: November 2016). Das Korpus besteht aus fünf genrespezifischen Subkorpora (*press*, *literature*, *science*, *official*, *personal*) sowie aus weiteren fünf regionalsprachlichen Subkorpora (*Hungary*, *Slovakia*, *Subcarpathia*, *Transylvania*, *Vojvodina*), die nach der einmaligen Registrierung zur Verfügung stehen. Es ist komplett morphosyntaktisch annotiert hinsichtlich Wortstamm, POS-Tag und Flexionsform.

---

4 www.sfs.uni-tuebingen.de/GermaNet (Stand: 18.11.2016).

5 https://wordnet.princeton.edu (Stand: 18.11.2016).

6 http://cvc.instituto-camoes.pt/tecnologias-da-lingua/temanet.html (Stand: 18.11.2016).

7 www.ids-mannheim.de/cosmas2/web-app (Stand: 18.11.2016).

8 www.cis.uni-muenchen.de/~schmid/tools/TreeTagger (Stand: 18.11.2016).

9 www.connexor.com/nlplib/?q=mpt (Stand: 18.11.2016).

10 http://corpus.nytud.hu/mnsz/index_eng.html (Stand: 18.11.2016).

INSTITUT FÜR DEUTSCHE SPRACHE

COSMAS II ▾ ■ Abmeldung ★ Recherche ⇄ Optionen Hilfe

Aktuelles Archiv: W - Archiv der geschriebenen Sprache
Aktuelle Suchanfrage:
Treffer:
Aktuelles Korpus:
Referenz: Deutsches Referenzkorpus (DeReKo-2015-II)
Aktive Treffer:

Archive | Korpus | Such. | Wortform. | Ergebnisse | Kook. | KWIC | Volltext | Export

Öffnen Sie bitte eines der folgenden Archive:

vordefinierte Archive
W - Archiv der geschriebenen Sprache
W2 - Archiv der geschriebenen Sprache
W3 - Archiv der geschriebenen Sprache
W4 - Archiv der geschriebenen Sprache
W-ÜBRIG - Archiv der aussortierten geschriebenen Korpora
PARONYM - Archiv des Paronym-Projekts des IDS
WP - Wikipedia Artikel u. Artikeldiskussion & Benutzerdiskussion 2013/2015
WPE - englischsprachige Wikipedia Artikel u. Artikeldiskussion & Benutzerdiskussion 2015
WP_FS - fremdsprachige Wikipedia Artikel u. Artikeldiskussion & Benutzerdiskussion 2015
TAGGED-C - Archiv morphosyntakt. annotierter Korpora (CONNEXOR)
TAGGED-C2 - Archiv morphosyntakt. annotierter Korpora (CONNEXOR)
TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger)
TAGGED-T2 - Archiv morphosyntakt. annotierter Korpora (TreeTagger)
TAGGED-M - Archiv morphosyntakt. annotierter Korpora (MECOLB; ehem. TAGGED)
HIST - Archiv der historischen Korpora
UMB - Archiv des Projekts "Sprachliche Umbrüche des 20. Jahrhunderts"
SNA - Projektkorpus Sprache 1933-1945
GFDS - Kartei der Gesellschaft für deutsche Sprache
SZ - Archiv der Süddeutschen Zeitung
WK-PH - Archiv der phasengegliederten Wendekorpora

Öffnen Sie bitte eines der folgenden Archive: Informationen zu den Archiven finden Sie unter Organisation des Textmaterials.

Impressum | Datenschutz | © 2003 - 2016 IDS Mannheim, COSMAS II ,,,,,, Version 2.2

Abb. 4: Startseite von COSMAS II

MNSZ2

Corpus: MNSZ2 Help

Search
Save
View options
KWIC/Sentence
Sort
Left | Right
Node
References
Shuffle
Sample
Filter
Frequency
Node forms
Node lemmas
Node analyses
Doc IDs
Collocations
Feedback

Corpus: MNSZ2
Hits: 67745

Page 1 of 3388 Go Next | Last

doc#3	Ugyanakkor a legmodernebb technológiákkal szerelte fel azt .
doc#85	Szemléletváltás történt a mezőgazdaságban , visszaszorulóban a magasabb terméshozam érdekében korábban bevetett agresszív módszerek , technológiák , nő a vadállomány élettere .
doc#90	Tökéletesítette a technológiát : az altárnák munkálatainak gyorsítása érdekében bevezette a sűrített levegővel hajtott fúrógépek alkalmazását ; javított a bányatérképezés módszerein , előmozdította a kohóüzemek tökéletesítését .
doc#91	A bányászat így vált Európában s aztán Magyarországon is igen korán a műszaki-technikai haladás úttörőjévé , s a szorosan vett szakmai fejlesztésen túl - mint Zsámboki László is hangsúlyozza - a gépészet , a mechanika , a geodézia , az ásványtan , a földtan , a kémia , a kémiai technológia , az erdészet stb. kutatásának műhelyévé is .
doc#91	1862-ben a szomolnai kohó számára új kénnyerési technológiát dolgozott ki .
doc#91	Tökéletesítette a technológiát : az altárnák munkálatainak gyorsítása érdekében bevezette a sűrített levegővel hajtott fúrógépek alkalmazását ; javított a bányatérképezés módszerein , előmozdította a kohóüzemek tökéletesítését .
doc#91	Ilyen többek közt A műszaki vegytan egyes fejezetei , melyet a " pénzügyi közegek , a gazdasági és iparintézetek számára " írt (Budapest , 1887) , a Szeszgyártás és ecetgyártás (Budapest , 1890) , a Kémiai technológia (Hankó Vilmos társszerzővel . Budapest , 1910) .
doc#91	Több tantárgyat is előadott , például a fakitermeléstant , a fa technológiáját , a vadászat- és fegyverismerettant stb.
doc#91	1910-ben újra visszatért Selmecbányára , ahol a fák kémiai technológiájával foglalkozott .
doc#98	Az írásra alkalmas új anyag előállításának technológiája csak a 10. században jutott el Európába az arabok által .
doc#98	Tóth István a sörgyártás technológiáját is részletesen leírta .
doc#186	Egy ilyen halálos fegyver előállítása , mint ezé a rugós szipkáé elég bonyolult technológiát feltételezett , mélyfagyasztó sem volt akkor még a háztartásokban , Kolozsvárt mifelénk például még a jegesember járt , de mi bíztunk tulajdon leleményességünkben , abban , hogy egy kis szorgalommal , esetleg mérnöki segítséggel a halálos fegyver kivitelezhető .
doc#253	De ugyanekkor oly társadalomban élünk , melynek nincs szellemi rendje , se stabilitása , nincsen általánosan elfogadott kozmológiája , ahol az orvostudomány és a higiénia fejlődése lekúzdhetetlen túlszaporodáshoz vezet és ahol a technológia minden előnye együttal megmérgezi a levegőt , a vizeket , a földet , és néhány évtizeden belül lakhatatlan sivataggá változtatja a világot .
doc#256	Miért jött így , kérdem . Talán azért , mert mindenki hanyatt feküdt a technológia elé s azt nézi csak , mit ad , s villanyborotva , diktafón , kompjúter , degeszig teli hűtőszekrény s tévé nélkül a létet el sem képzeli ?
doc#256	Tanuld meg ezt a versemet , hogy elkísérjelek . Lehet , s túléled még az ezredet , s pár kurta évre kiderül , mert a bacillusok dühödt revánsa mégse sikerül , s a technológia mohó hadosztályai több erőt mozgatnak , mint a földgolyó - memóriádból szedd elő s dúdold el még egyszer velem e sorokat : mert hova lett a szépség és a szerelem ?
doc#279	Nálunk általában kétféle könyv került fordítói kezünkbe : technikai , műszaki ismereteket , új technológiákat magyarázó szakkönyvek , továbbá legálisan ki nem adható politikatudományi művek , melyek a nyugati világ jobb megértését voltak hivatottak előségiteni az érdeklődő (bár számunkra ismeretlen) politikusi olvasók számára .
doc#331	Új építészet keletkezett , új technológia ; fogalmam sincs , mihez fogtak volna , ha azok az amerikai konzervdobozok nincsenek .
doc#341	Tessék mondani , mért kell nekem tudni a plasztikbomba technológiáját - amikor nyugdíj előtt állok ?
doc#347	A technológia csodája volt ez a hiperháború ; s ennek végeztével folytatódik tovább a kevésbé látványos emberirtás a blokáddal : azóta is tömegesen hullanak el a gyerekek Irakban .
doc#347	Ami az újabb technológiából fog következni , azt már úgysem érem fel gyatrán betáplált , elmaradt képzeletemmel .

Page 1 of 3388 Go Next | Last

Lexical Computing Ltd.
Sketch Engine (ver:open-2.59.1-open-2.91.13)
A member of the Hungarian Corpus Portal.
MNSZ2 · v2.0.4 · 2016-08-29
Interface language: English | magyar

Abb. 5: Ergebnis für „technológiákkal“ im HNC

Im Bereich der Wörterbuchdokumentation gilt das von der Oxford University Press herausgegebene *Oxford English Dictionary* (OED) als einer der größten Lieferanten lexikalischen Inhalts der englischen Sprache. Das Wörterbuch, auf das man kostenlos online zugreifen kann, liefert sowohl die aktuelle(n) wie auch die historische(n) Bedeutung(en) von ca. 600.000 Wörtern. Diese stützen sich auf nahezu 3 Millionen Belege aus spezifischen Epochen der Genres klassischer Literatur, Filmskripten und Kochbüchern.

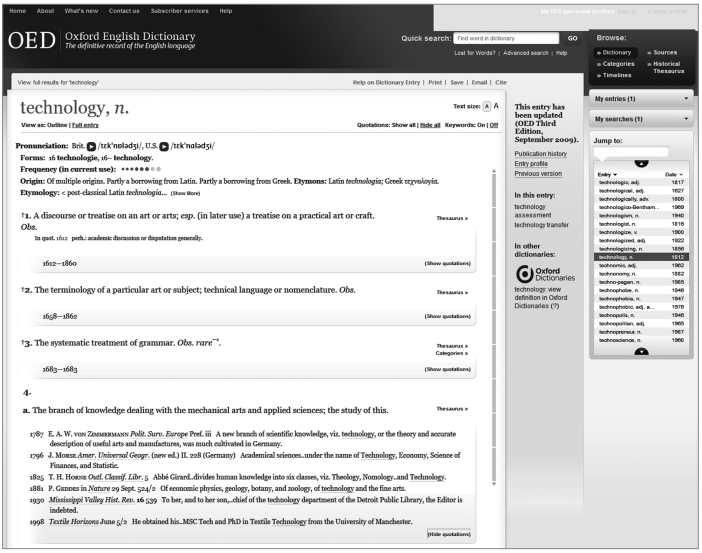


Abb. 6: Ergebnis für „technology“ im OED

4. Notwendigkeit einer Forschungsinfrastruktur

Die Austauschbarkeit von Sprachdaten und -technologien durch eine ausgeprägte Infrastruktur innerhalb der Forschungslandschaft ist nicht nur ein Hauptziel der EFNIL. Der europäische Forschungsverbund *Common Language Resources and Technology Infrastructure* (CLARIN) hat ebenjenes Ziel, eine stabile Forschungsinfrastruktur aufzubauen und zur Verfügung zu stellen, das dem interessierten Benutzer erlaubt, geschriebene, gesprochene sowie multimodale Sprachdaten aufzufinden, zu verwenden und mit geeigneten Tools weiter zu verarbeiten. So ermöglicht der von CLARIN bereitgestellte Webservice *Virtual Language Observatory*[11] das Suchen nach Ressourcen, während es die *Federated Content Search*[12] erlaubt, in Ressourcen zu suchen. Darüber hin-

[11] https://vlo.clarin.eu (Stand: 18.11.2016).

[12] http://weblicht.sfs.uni-tuebingen.de/Aggregator (Stand: 18.11.2016).

aus können die Ressourcen durch *WebLicht*[13] mit den zur Verfügung stehenden Sprachanalysetools verarbeitet werden. Um dies zu ermöglichen, sind alle Ressourcen standardisiert und mit Metadaten versehen.

Die Forschungsinfrastruktur wird derzeit von 38 CLARIN-Zentren gepflegt und weiter ausgebaut, wobei auch externe Wissenschaftler und Institute einen Webservice bereitstellen können. Das Institut für Deutsche Sprache bildet eines dieser Zentren. Das Clarin-D Projekt (vormals D-Spin), das bis März 2011 vom Bundesministerium für Bildung und Forschung (BMBF) gefördert wurde, hat die Aufgabe, diese Maxime der CLARIN-Initiative innerhalb Deutschlands zu gewährleisten.

Den institutionellen Rahmen für den europaweiten Zusammenschluss bildet das CLARIN-ERIC, dessen neun Gründungsmitglieder Bulgarien, Estland, Dänemark, Deutschland, Niederlande, Österreich, Polen, die Tschechische Republik sowie die Niederländische Sprachunion sind. Seit August 2016 ist auch Ungarn in CLARIN-ERIC als 19. Mitglied vertreten.

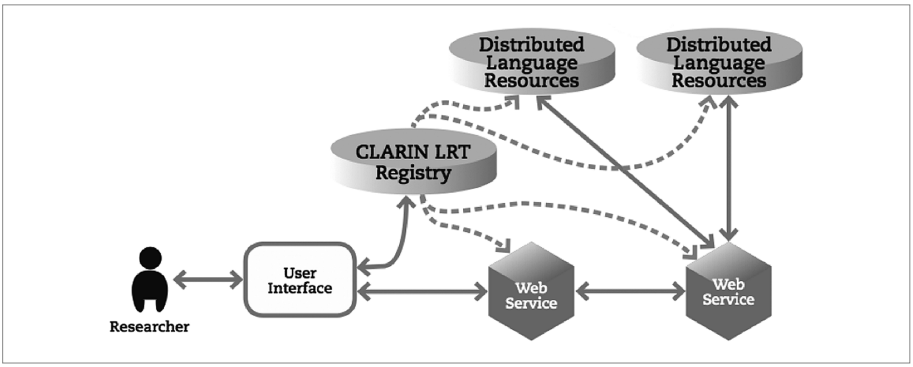


Abb. 7: Schematische Darstellung der CLARIN-Infrastruktur

Eine weitere Plattform zum Austausch von digitalen Sprachressourcen bietet die *Multilingual Europe Technical Alliance* (META-NET),[14] die neben Wissenschaftler auch Technologieanbieter und -nutzer vereint. Ziel dieser Vereinigung, an der sich derzeit 34 Nationen – darunter auch Ungarn mit der Ungarischen Akademie – beteiligen, ist die Entwicklung einer Infrastruktur zum Austausch von digitalen sprachtechnologischen Ressourcen durch die Vernetzung verschiedener Technologiefelder. Für den Zugriff auf die Sprachressourcen samt Metadaten wurde die sich stetig weiterentwickelte Plattform

[13] http://weblicht.sfs.uni-tuebingen.de (Stand: 18.11.2016).

[14] www.meta-net.eu (Stand: 18.11.2016).

META-SHARE,[15] die gleichsam kostenlose wie auch kostenpflichtige Ressourcen aufnimmt, entworfen. Um einen stabilen Austausch zu gewährleisten, werden alle Ressourcen standardisiert.

5. EFNIL-Mitglieder (*Umfrage*)

Die Anzahl der EFNIL-Mitglieder ist seit seiner Gründung von 12 auf 27 Nationen mit zusammen 40 beteiligten Instituten gestiegen. Jedes Institut beobachtet und analysiert die in ihrer Nation anerkannten Sprachen und stellt die Sprachressourcen unterschiedlicher Menge und Art nach Möglichkeit zur Verfügung. Die Menge an verfügbaren digitalen Sprachressourcen nimmt nicht zuletzt aufgrund der Nachfrage stetig zu. Eine stärkere Transparenz nicht nur hinsichtlich der verfügbaren Sprachressourcen sondern bezüglich der IT-Infrastruktur, aber auch dem Wissen um die Zielgruppe verstärkt den Austausch zwischen den EFNIL-Mitglieder und fördert somit die Gewährleistung guter Forschungsarbeit.

Um die Ziele der EFNIL weiter auszubauen und zu fördern, wurde unter anderem in Erfahrung gebracht, welche Ressourcen die einzelnen Institute der EFNIL-Mitglieder bereitstellen (können), wie sie technisch aufgestellt sind und wie groß die Nachfrage nach den individuellen Sprachdaten ist.[16]

Jedes der EFNIL-Mitglieder stellt eine unterschiedliche Bandbreite an digitalen Sprachressourcen zur Verfügung. Viele Nationen haben eigene Korpora, z.B. das *Bulgarian National Corpus* (BulNC), das *Czech National Corpus* (CNC), das *Hungarian National Corpus* (HNC), das *TLIO Textual Corpus*, das *Corpus de la Parole* oder das *Slovak National Corpus*, um nur ein paar zu nennen. Größtenteils handelt es sich um schriftsprachliche (Parallel-)Korpora, es existieren aber auch gesprochensprachliche sowie multimodale. Die Größe der Korpora variiert dabei von mehreren Millionen bis hin zu mehreren Billionen Wörtern. Darüber hinaus stehen in einigen Korpora sprachverarbeitende Tools zur weiteren Verarbeitung zur Verfügung, wie POS-Tagger, Lemmatizierer, Chunker, usw. Daneben existieren einige Wörterbücher – von (historischen) Rechtschreib- bis hin zu Neologismus-Wörterbücher –, wie z.B. das *Oxford English Dictionary*, sowie Datenbanken, welche die lexikalisch-semantische Relation zwischen den Wörtern darstellt, wie z.B. das *FranceTerme* oder das *Bulgarian Wordnet*. Der Zugriff auf diese digitalen Ressourcen ist teilweise beschränkt oder kostenpflichtig. Einige dieser digitalen Ressourcen sind auch als bzw. über Apps erhältlich.

---

[15] www.meta-share.org (Stand: 18.11.2016).

[16] Aus Datenschutzgründen können einige Institute keine Auskunft über die technische Ausstattung oder Nutzerinformationen geben.

Verwaltung, Pflege und Ausbau dieser Datenmengen benötigen ein geeignetes Rechenzentrum. Jedoch besitzen lediglich zwei von acht Nationen ein eigenes, während sechs Mitglieder ihre Daten ganz oder teilweise über ein externes Rechenzentrum managen. Teilweise ist die Auslagerung auf Kooperationsprojekte zurückzuführen, bei denen die Daten beim Kooperationspartner und somit auswertig verwaltet werden, teilweise ist die Art der Verwaltung aber auch abhängig von der Art der digitalen Ressource. So gab ein Mitglied an, dass multimodale Sprachressourcen, wie (Lern- und Lehr-) Videos, über einen externen Anbieter gehostet werden. Die digitalen Ressourcen werden dabei teilweise über Windows- teils über Linuxserver gemanaged; als Datenbank wird Oracle präferiert. Über die verfügbare Speicherkapazität ist wenig bekannt; lediglich ein Mitglied gab an, bis zu 90 TB bzw. 280 TB auf den Servern zur Verfügung zu haben. Die Websites, über die auf die jeweilige Ressource zugegriffen werden kann, werden über unterschiedliche, teils nationenspezifische Content-Management-Systeme gepflegt.

Bei der Frage nach dem Wissen über die Benutzer ihrer Ressourcen konnten 10 von 14 EFNIL-Mitgliedern ihre Zielgruppe definieren und somit ihr Angebot an Sprachressourcen und Tools dementsprechend anpassen und erweitern. Wie oft welche digitale Ressource und von wie vielen Benutzern sie jährlich verwendet wird, unterscheidet sich je nach Institution und Ressource; von einigen tausend bis hin zu mehreren hunderttausend Besuchern pro Jahr, sodass mehrere Millionen Aufrufe verzeichnet werden können. Über die Benutzergruppe ist bekannt, dass es sich hauptsächlich um Wissenschaftler oder um Lerner sowie Lehrer der jeweiligen Sprache handelt. Darüber hinaus werden die frei zugänglichen Ressourcen von Dritten weiterverwendet; während manche Ressource in Apps verarbeitet wird, findet sich eine im Rahmen eines TV-Quiz wieder.

6. Ausblick

Es wird angestrebt, sämtliche Informationen über alle vorhandenen Sprachressourcen inklusive deren Zugriffsrechte systematisch in EFNIL aufzuarbeiten und kontinuierlich bei Bedarf zu aktualisieren. Ebenso wird eine Transparenz hinsichtlich der informationstechnologischen Infrastruktur befürwortet, um den Austausch und die Vernetzung der Mitglieder zu fördern.

7. Literatur

COSMAS II – Corpus Search, Management and Analysis System. Das Portal für die Korpusrecherche in den Textkorpora des Instituts für Deutsche Sprache. www.ids-mannheim.de/cosmas2 (Stand: September 2016).

Grimm, Jacob/Grimm, Wilhelm (1854): Deutsches Wörterbuch. Leipzig.

Francis, W. Nelson/Kučera, Henry (1979): Brown Corpus Manual – Manual of information to accompany a Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Brown university. http://clu.uni.no/icame/brown/bcm.html#bc1 (Stand: Januar 2017).

Murray, James et al. (1884): A new English dictionary on historical principles. Oxford.

Teubert, Wolfgang/Belica, Cyril (2014): Von der linguistischen Datenverarbeitung am IDS zur „Mannheimer Schule der Korpuslinguistik". In: Institut für Deutsche Sprache (Hg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Red.: Melanie Steinle, Franz Josef Berens. Mannheim, S. 298-319.