EUROPEAN LANGUAGE EQUALITY

D1.10

Report on the Dutch Language

Authors	Frieda Steurs, Vincent Vandeghinste, Walter Daelemans
Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.10
Deliverable title	Report on the Dutch Language
Type	Report
Number of pages	23
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Frieda Steurs, Vincent Vandeghinste, Walter Daelemans
Reviewers	Stefanie Hegele, Ainara Estarrona
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland
	Prof. Dr. Andy Way – andy.way@adaptcentre.ie
	European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany
	Prof. Dr. Georg Rehm – georg.rehm@dfki.de
	http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
CALL	Computer-assisted language learning
CGN	Corpus Gesproken Nederlands (Corpus Spoken Dutch)
CL	Computational Lingusitics
CLARIN	Common Language Resources and Technology Infrastructure
CLIN	Computational Linguistics in the Netherlands
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DPC	Dutch Parallel Corpus
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
ELDA	European Language Distribution Agency
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
EU	European Union
FWO	Fonds voor Wetenschappelijk Onderzoek (Fund for Scientific Research)
GiGaNT	Groot Geïntegreerd lexicon van de Nederlandse Taal (Large Integrated lexicon of the Dutch Language)
GDPR	General Data Protection Regulation
GPU	Graphical Processing Unit
HPC	High-Performance Computing
INT	Instituut voor de Nederlandse Taal (Dutch Language Institute)
LDC	Linguistic Data Consortium
LR	Language Resources/Resources
LT	Language Technology/Technologies
ML	Machine Learning
NER	Named Entity Recognition
NL	the Netherlands
NLG	Natural Language Generation
NLP	Natural Language Processing
NWO	Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Dutch Organisation for Scientific Research)
NOTaS	Nederlandse Organisatie voor Taal- en Spraaktechnologie (Dutch Organisation for Language and Speech Technology)

O	Object
OCW	Onderwijs, Cultuur en Wetenschap (Education, Culture and Sciences)
ODWN	Open Dutch WordNet
OPUS	Open Parallel Corpus
POS	Part-of-speech
S	Subject
SoNaR	Stevin Nederlands Referentiecorpus (Stevin Dutch Reference corpus)
SOV	Subject-object-verb
STEVIN	Spraak- en Taaltechnolgische Voorzieningen in het Nederlands (Speech and Language Technological Resources for Dutch)
SVO	Subject-verb-object
UD	Universal Dependencies
V	Verb
W3C	World Wide Web Consortium

Abstract

This report provides a new state of affairs with regard to language technology for Dutch, a language with approx. 25 million speakers. Language technology for Dutch is highly developed and the importance and status of Dutch is confirmed by other measurements, such as the number of online sales, which is growing strongly, and so is the presence of Dutch online.

The Netherlands and Flanders have a strong research community in speech and language technology, which means that a lot of digital data is available for Dutch, such as databases for spoken and written language, associated software for the analysis and processing of language data, etc. Many of these language materials are freely accessible to researchers (via CLARIN), and often also to everyone via the Dutch Language Institute (INT). Companies can therefore easily include Dutch as one of the available languages, when developing new digital applications. Internationally, the Dutch language is well represented in the major European networks, such as CLARIN, ELRC, ELG, ELE and ELRA.

A lot of language data is available, including via https://taalmaterialen.ivdnt.org. We note a number of important corpora such as the SoNaR reference corpus, the Dutch Parallel Corpus (DPC) and the Corpus Spoken Dutch (CGN). There is an urgent need for updated versions of these corpora with more, more recent and diversified language data. There is also a need for new corpora with, for example, the language of social media.

In terms of lexical data, the INT is currently developing a computational lexicon for Dutch from the sixth century to contemporary Dutch with two major parts: a historical and a modern lexicon. We also mention Open Dutch Wordnet (ODWN), a semantic database with limited content. A new and more extensive version of this is also desirable.

There are many software applications available for text and speech analysis and processing for Dutch. In translation technology, too, Dutch is well represented in the most well-known translation software packages, both for computer-aided translation and for machine translation.

In order to keep up with the latest developments in the field of digital language infrastructure, various organisations and working groups are active in Belgium and the Netherlands. We mention the Dutch AI coalition and the NL Speech coalition where (computational) linguists and other experts meet to discuss the latest trends in machine learning and artificial intelligence. The Dutch NOTaS foundation represents research institutions and application developers in the Language and Speech Technology sector. A network of computational linguists and researchers is formed by CLIN (Computational Linguistics in the Netherlands) and organises conferences alternately in Belgium and the Netherlands. Other consultation platforms in Belgium include Belgium NLP meetup, which brings researchers and companies together around NLP, and Flanders AI, the research program for AI in Flanders.

Many of the tools and materials mentioned were developed in the STEVIN research program (2004-2011) in which the Netherlands (NWO) and Flanders (FWO) joined forces. The need for a new joint research program is great. Through joint projects, Flanders and the Netherlands can act more vigorously and tap into new lines of research.

Samenvatting

Dit rapport geeft een nieuwe stand van zaken wat betreft de taaltechnologie voor het Nederlands. Er zijn ongeveer 25 miljoen Nederlandstaligen, waarvan 17 miljoen in Nederland wonen, 6,5 miljoen in België, en 400.000 in Suriname. Daarmee is het Nederlands een van de 40 meest gesproken talen in de wereld. Het Nederlands alleen wordt door meer mensen gesproken dan de Noord-Germaanse (Scandinavische) talen bij elkaar. Het is de achtste taal in de Europese Unie, de twaalfde taal op internet en een belangrijke taal in de sociale media.

De Nederlandstalige Wikipedia staat op plaats zes van de wereld. De taaltechnologie voor het Nederlands is hoog ontwikkeld. Ook andere metingen bevestigen het belang en de status van het Nederlands: met name de aanwezigheid van het Nederlands als taal op internationale websites en het surfgedrag van Nederlanders en Vlamingen. Nederland behoort tot de kopgroep van de EU-28-landen met de meeste gezinnen die toegang hebben tot internet en ook België scoort bovengemiddeld.

Nederland en Vlaanderen zijn economisch sterk en dat vertaalt zich in vele websites van bedrijven waar telkens het Nederlands als taal aanwezig is. De online verkoop groeit sterk en dus ook de aanwezigheid van het Nederlands online. Nederland en Vlaanderen beschikken over een sterke onderzoeksgemeenschap in de spraak- en taaltechnologie en daardoor zijn er veel digitale data beschikbaar voor het Nederlands, zoals databanken voor gesproken en geschreven taal, bijhorende software voor de analyse en verwerking van taaldata etc. Veel van deze taalmaterialen zijn vrij toegankelijk voor onderzoekers (via CLARIN), en vaak ook voor iedereen via het Instituut voor de Nederlandse Taal (INT). Zo kunnen bedrijven bij het ontwikkelen van nieuwe digitale toepassingen ook het Nederlands meenemen als een van de beschikbare talen. Internationaal is de Nederlandse taal goed vertegenwoordigd in de grote Europese netwerken, zoals CLARIN, ELRC, de European Language Grid, ELE en ELRA.

Er zijn veel taaldata beschikbaar, onder meer via https://taalmaterialen.ivdnt.org. We noteren een aantal belangrijke corpora zoals het SoNaR referentiecorpus, het Dutch Parallel Corpus (DPC) en het Corpus Gesproken Nederlands (CGN).Er is dringend behoefte aan actuele versies van deze corpora met meer, meer recente en gediversifieerde taaldata. Er is ook nood aan nieuwe corpora met bijvoorbeeld de taal van de sociale media.

Qua lexicale data ontwikkelt het INT momenteel een computationeel lexicon voor het Nederlands van de zesde eeuw tot het hedendaagse Nederlands met twee grote onderdelen: een historisch en een modern lexicon. We vermelden ook Open Dutch Wordnet, een semantische databank met beperkte inhoud. Een nieuwe en meer uitgebreide versie hiervan is eveneens wenselijk.

Er zijn heel wat software toepassingen beschikbaar voor tekst- en spraakanalyse en -bewerking voor het Nederlands. Ook in de vertaaltechnologie is het Nederlands goed vertegenwoordigd in de meest bekende vertaalsoftwarepakketten, dit zowel voor computerondersteund vertalen als voor automatische vertaling. Zowel in Google Translate als DeepL, maar ook in e-Translation is het Nederlands beschikbaar.

Om bij te blijven met de nieuwste ontwikkelingen op het vlak van digitale taalinfrastructuur zijn er verschillende organisaties en werkgroepen actief in België en Nederland. We noemen hier de Nederlandse AI-coalitie en de NL Spraak Coalitie waar (computationele) taalkundigen en andere experten mekaar ontmoeten om de nieuwste trends in machine learning en artificiële intelligentie te bespreken. De Nederlandse stichting NOTaS vertegenwoordigt onderzoeksinstellingen en applicatieontwikkelaars in de taal- en spraaktechnologiesector. Een netwerk van computationeel taalkundigen en onderzoekers wordt gevormd door CLIN (Computational Linguisitcs in the Netherlands) en organiseert afwisselend in België en Nederland congressen. Andere overlegplatformen in België zijn ondermeer Belgium NLP meetup, die onderzoekers en bedrijven samenbrengt rond NLP, en Flanders AI, het onderzoeksprogramma voor AI in Vlaanderen.

Veel van de vermelde tools en materialen werden ontwikkeld in het STEVIN onderzoeksprogramma (2004-2011) waarbij de krachten gebundeld werden tussen Nederland (NW0) en Vlaanderen (FWO). De nood aan een nieuw gemeenschappelijk onderzoeksprogramma is groot. Via gemeenschappelijke projecten kunnen Vlaanderen en Nederland krachtiger optreden en nieuwe onderzoekslijnen aanboren.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and it seeks to not only delineate the current state of affairs for each of the European languages covered in this series, but to additionally – and most importantly – to identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

To this end, more than 40 research partners, experts in more than 30 European languages have conducted an enormous and exhaustive data collection procedure that provided a detailed, empirical and dynamic map of technology support for our languages.[1]

This report has been developed by the European Language Equality (ELE) project.[2] With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the ELE project develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

2 The Dutch Language in the Digital Age

2.1 General Facts

Dutch is a West-Germanic language spoken by about 25 million people as a first language and 5 million people as a second language, i. e. most of the population of the Netherlands (where it is the only official language countrywide) and about 60% of the population of Belgium, mainly in Flanders (as one of three official languages)(Steurs, 2021). It is the third most widely spoken Germanic language, after its close relatives English and German.

This report focuses on the Dutch language and LT for the Netherlands and Flanders, the so called Low Countries. Outside the Low Countries, it is the native language of the majority of the population of Suriname where it also holds an official status, as it does in Aruba, Curaçao and Sint Maarten which are located in the Caribbean. Historical linguistic minorities on the verge of extinction remain in parts of France and Germany, and in Indonesia (Java and Bali), while up to half a million native speakers reside in the United States, Canada and Australia combined.

The Ministry of OCW (Education, Culture and Sciences) organises and monitors education in general, including the education of the Dutch language in the Netherlands. In Flanders, the Department Onderwijs & Vorming (Department of Education and Training) is responsible for education.

Language skills are the key qualification needed in education as well as for personal and professional communication. The education of Dutch *extra muros* is also systematically monitored via studies performed by or under the supervision of the Dutch Language Union.[3] They also issue concrete policy and practical guidelines for addressing problems in areas such as spelling, reading skills, language competence of teachers, language and/or educational retardation, education in literature, and others.[4] Continuous attention to Dutch language teaching in schools is essential for providing students with the language skills required

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.
[2] https://european-language-equality.eu
[3] https://taalunie.org
[4] https://taalunie.org/publicaties/189/meerjarenbeleidsplan-2020-2024

for an active participation in society. Language technology makes an important contribution here by offering computer-assisted language learning (CALL) systems, and educational games, which allow students to experience language in a playful way, for example by linking special vocabulary in electronic text to comprehensible definitions or to audio or video files supplying additional information. Search and text categorisation methods can also help in finding exercises and texts suited for specific language proficiency levels.

Dutch is one of the closest relatives of both German and English. Dutch, like English, has not undergone the High German consonant shift and does not use Germanic umlaut as a grammatical marker, has largely abandoned the use of the subjunctive, and has lost much of its morphology, including most of its case system. Features shared with German include the survival of two to three grammatical genders, as well as the use of modal particles, final-obstruent devoicing, and a similar word order. Dutch vocabulary is mostly Germanic and incorporates slightly more Romance loans than German but far fewer than English.[5]

Certain linguistic characteristics of Dutch are challenging for computational processing. The Dutch language exhibits some specific characteristics, which contribute to the richness of the language by allowing the speakers to express ideas in a large variety of ways. One such particularity is that it is quite common to put non-subjects sentence-initially (much more common than in English). For example, consider the English sentence *the woman was going to the store every day*. In English, there are very limited possibilities to use a different word order in this sentence, but in the Dutch equivalent almost any phrase can be the initial phrase in the sentence:

De vrouw ging elke dag naar de winkel.
Elke dag ging de vrouw naar de winkel.
Naar de winkel ging de vrouw elke dag.

Word order in Dutch is thus much freer than in English (but not as free as in German). For the main clause, Subject Verb Object (SVO) word order applies; for example:

Pieter (S) eet (V) een appel (O).

whereas the subclause has the SOV wordorder, for example:

(Ik weet) dat Pieter (S) een appel (O) eet (V).

These flexible word order patterns may cause issues in NLP that do not occur in English.

Also, the Dutch language is quite productive in creating new compounds, though the use and productivity of compounding is not as extreme as in German. Nevertheless, newly formed compounds occur frequently and are difficult to process for NLP technology. Another characteristics of Dutch that makes processing difficult is formed by separable verb prefixes that can occur far from the verb in nested constructions like:

Hij stelde zich na mij een drankje aangeboden te hebben en wij in gesprek geraakt waren aan ons voor. (He introduced himself after he offered me a drink and we started a conversation.)

The meaning of a verb containing such a separable prefix like *voor*, *in* or *uit* can very often not be derived from the meaning of the base verb and the meaning of the prefix.

Dutch has a variety of dialects that are described in dialect dictionaries and databases. Dutch is a monocentric language, at least what concerns its written form, with all speakers using the same standard form (authorised by the Dutch Language Union) based on a Dutch orthography defined in the so-called "Green Booklet" authoritative dictionary[6] and employing the Latin alphabet when writing. The standard is obligatory in education and governmental publications. There is lexical variety between dialects, but also between the standard

[5] https://www.ethnologue.com/language/nld
[6] https://woordenlijst.org

language in the Netherlands and Flanders. These varieties are described in dictionaries and different text corpora prove the existence and contextual use. There are also divergences in the pronunciation between the Netherlands and Flanders. In contrast to its written uniformity, Dutch lacks a unique prestige dialect and has a large dialectal continuum consisting of 28 main dialects, which can themselves be further divided into at least 600 distinguishable varieties. In the Netherlands, the Hollandic dialect dominates in national broadcast media while in Flanders Brabantian dialect dominates in that capacity, making them in turn unofficial prestige dialects in their respective countries.

The Netherlands and Belgium produce the vast majority of music, films, books and other media written or spoken in Dutch. Foreign films and television series are subtitled.

2.2 Dutch in the Digital Sphere

The Low Countries are a very rich and economically active region, and this translates into a lot of websites with Dutch as a language. In 2020, there were 6,109,589 websites registered with .nl extension, and 1,605,288 with .be extension. Apart from that, a lot of .com and .org sites include Dutch. According to the W3C 0.55% of all websites (1.18 billion) are in Dutch.[7] Due to the rise of e-commerce, a lot of webshops also include Dutch.

In the Netherlands, 96% of the population is an internet user;[8] in Flanders 93% of the population is an internet user.[9] In 2021 the Dutch Wikipedia is the sixth-largest Wikipedia edition, with 2,070,744 articles.

Over the last ten years we have seen some interesting developments since the Dutch language is very much used on social media (Twitter, Facebook, WhatsApp, Instagram, Youtube etc.).[10] Nearly three million people used Twitter in the Netherlands in 2021, an increase of over 100,000 users compared to the previous year.[11] The same trends can be seen in Belgium. The growth on all platforms is significant and still increasing. This leads to the development of new linguistic trends and sublanguages that can be studied and analysed. At the same time, we can see a growth in new linguistic studies into language variation, with attention for new language data and corpus material such as the language of some youth groups (Morrocorp),[12] streetlanguage, slang, the language of youth in areas with a lot of migration (Marzo, 2017), etc.

3 What is Language Technology?

Natural language[13] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialised field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, Language Technology (LT). While there are differences in focus and orientation,

[7] https://w3techs.com/technologies/details/cl-nl-
[8] https://mindwize.nl/blogs/digitaal-gebruik-nederland-2021/
[9] https://www.statistiekvlaanderen.be/nl/internetgebruik-naar-gebruiksfrequentie
[10] https://www.coosto.com/nl/blogs/social-media-gebruik-2021-cijfers-statistieken
[11] https://www.statista.com/statistics/880865/number-of-twitter-users-in-the-netherlands/
[12] https://taalmaterialen.ivdnt.org/download/tstc-moroccorp-2/
[13] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of Sections 1 and 2 of Aldabe et al. (2021).

since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably AI), mathematics and psychology among others. In practice, these communities work closely together, combining methods and approaches inspired by both, together making up *language-centric AI*.

Language Technology is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, whether they are written, spoken or embodied.

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning, rule-based systems have been displaced by data-based ones, i. e. systems that learn implicitly from examples. In the recent decade of 2010s we observed a radical technological change in NLP: the use of multilayer neural networks able to solve various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

In recent years, the LT community has been witnessing the emergence of powerful new deep learning techniques and tools that are revolutionising the way in which LT tasks are approached. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The success in these areas of AI has been possible because of the conjunction of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual data), 3) increase in high performance computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches.

LT is trying to provide solutions for the following main application areas:

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e. the generation of speech given a piece of text, Automatic Speech Recognition, i. e. the conversion of speech signal into text, and Speaker Recognition (SR).
- **Machine Translation**, i. e. the automatic translation from one natural language into another.
- **Information Extraction and Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query.

- **Natural Language Generation (NLG)**. NLG is the task of automatically generating texts. Summarisation, i. e. the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.
- **Human-Computer Interaction** which aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). A very popular application within this area are conversational agents (better known as chatbots).

LT is already fused in our everyday lives. As individual users we may be using it without even realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just very few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[14]

4 Language Technology for Dutch

There are several European infrastructures which include Dutch tools and resources, and which are offering these largely for free. We mention the European Language Grid[15], which is accessible to anyone, CLARIN[16], which is targeting academic users, and ELRC (European Language Resource Coordination)[17], an initiative from the European institutions, as the most prominent ones.

Alternatively there are the paid membership services of the European Language Distribution Agency (ELDA)[18] and its American counterpart Linguistic Data Consortium (LDC)[19] which also include several datasets containing Dutch.

We keep a detailed list of available tools and resources for Dutch at K-Dutch,[20] the CLARIN Knowledge Centre for Dutch at the Dutch Language Institute.

[14] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

[15] https://www.european-language-grid.eu

[16] https://www.clarin.eu

[17] https://www.lr-coordination.eu

[18] http://www.elra.info

[19] https://catalog.ldc.upenn.edu

[20] https://kdutch.ivdnt.org

4.1 Language Data

There are several corpora available for Dutch, for research as well as for commercial purposes. Many of these are downloadable from Taalmaterialen.[21]

The SoNaR corpus[22] (Oostdijk et al., 2013) is constructed as a reference corpus, containing material from different text genres, both from the Netherlands and Belgium. The research version contains 500 million words, whereas the commercial version contains 271 million words. About one million of these words have been manually annotated or corrected with syntactic (van Noord et al., 2013) and semantic information, and is available as a separate download.

A parallel corpus of 10 million words for the language pairs Dutch-English and Dutch-French is the Dutch Parallel Corpus (DPC)[23] (Paulussen et al., 2013). Dutch is also available in several parallel sub-corpora from OPUS, the Open Parallel Corpus,[24] (Tiedemann, 2012) and from ELRC[25] and the other already mentioned infrastructures in Section 4.

The Corpus Gesproken Nederlands (CGN) (Oostdijk et al., 2002) (Corpus Spoken Dutch)[26] is a collection of 900 hours (almost 9 million words) of contemporary spoken Dutch (1998-2004) from native speakers in Flanders and the Netherlands. The speech recordings are aligned with several transcriptions (e. g. orthographic, phonetic) and annotations (syntax, POS-tags). There is a large demand for a new large corpus for spoken Dutch containing more recent language and more variants, in order to train speech recognition engines.

In addition different research groups have developed and made available a diverse set of corpora for specific domains and applications; there is an open attitude to sharing corpora for research. However, for corpora with social media data, it becomes increasingly difficult to share and even collect the necessary material due to restrictions in the EU's GDPR. This seriously hampers research in this domain.

The Dutch Language Institute is developing a computational lexicon of the Dutch language from the sixth century up to the present (Ruitenberg et al., 2010). This lexicon, called GiGaNT, will be a collection of words and word groups, including named entities (names of persons, places, organisations), showing every possible variant of spelling and form.

The lexicon has two main modules: Hilex, the historical lexicon component which is available through a webservice, and Molex, the modern lexicon component, containing materials from the INT corpora.[27]

Open Dutch WordNet[28] (Postma et al., 2016) is a Dutch lexical semantic database which was created by removing the proprietary content from Cornetto,[29] and by using open source resources to replace this proprietary content. As a result, it has a rather limited coverage and is mainly focused on verbs and nouns. It is freely available. A version with more coverage and more contemporary words would be desirable.

Hugging Face[30] is a large open-source community that quickly became a hub for pre-trained deep learning models, mainly aimed at NLP. Their core mode of operation revolves around the use of Transformers. They lists 112 different publicly available pre-trained BERT-like (Devlin et al., 2019) language models for Dutch. Word2vec embeddings (Mikolov et al., 2013) for Dutch are available from Tulkens et al. (2016).[31] Nevertheless there is still demand

[21] https://taalmaterialen.ivdnt.org
[22] http://hdl.handle.net/10032/tm-a2-h5
[23] http://hdl.handle.net/10032/tm-a2-h3
[24] https://opus.nlpl.eu
[25] https://www.elrc-share.eu
[26] http://hdl.handle.net/10032/tm-a2-k6
[27] http://hdl.handle.net/10032/tm-a2-p9
[28] https://github.com/cltl/OpenDutchWordnet
[29] Cornetto is a lexical semantic database which is no longer distributed due to intellectual property reasons.
[30] https://huggingface.co
[31] https://github.com/clips/dutchembeddings

for very large scale language models for Dutch, and for language models on certain domains and registers.

4.2 Language Technologies and Tools

For linguistic text analysis, several tools are available for download or as online services: Frog[32] (van den Bosch et al., 2007) provides lemma, morphological segmentation, part of speech tagging, named entity type, base phrase chunk, and typed dependency information. The Alpino parser[33] (van Noord, 2006), which is a hybrid knowledge-based/statistical parser, provides deep linguistic dependency parsing. Pattern[34] and LeTs[35] are multilingual tools for text analysis including Dutch.

SpaCy,[36] an open source software library for advanced natural language processing contains Dutch models. The same holds for Stanza,[37] a collection of tools for linguistic analysis, and for UDPipe,[38] a trainable pipeline for tokenisation, tagging, lemmatisation and dependency parsing. Dutch NER is available in OpenNLP.[39] A UDPipe instantiation, providing POS tagging, lemmatisation and universal dependency parsing runs at Lindat.[40] The Weblicht service[41] CLARIN-D/SfS-Uni. Tübingen (2012) also provides analysis for Dutch, but is only available behind the CLARIN login.

Text-to-speech engines are available from commercial vendors, often with two language variants, one for Belgian Dutch (also known as Flemish) and one for Netherlandic Dutch.

Speech recognition is available from several commercial vendors, but there are also ASR engines for research purposes, for both variants of Dutch.

Dutch is present in most commercial online translation services, such as Google translate,[42] Bing[43] and DeepL,[44] which provide a limited amount of translation for free. eTranslation[45] from the European Commission provides unlimited translation, including from and to Dutch.

For Language Generation there are some models available on Hugging Face, such as GPT-2 models, for which there is also a demo.[46]

We are not aware of any specific tools for Dutch concerning Information Extraction, Information Retrieval nor Human Computer Interaction.

4.3 Projects, Initiatives, Stakeholders

There are several initiatives in the Netherlands and Belgium to group initiatives, disseminate achievements and join efforts, leading to a joint NLP community, but there is currently no overarching programme for the further development of tools and resources for Dutch. The language technology community in the Netherlands and Flanders would be very much in favour of setting up a follow-up programme to the STEVIN programme (Spyns and

[32] http://languagemachines.github.io/frog/
[33] https://github.com/rug-compling/alpino-docker
[34] https://github.com/clips/pattern
[35] https://lt3.ugent.be/resources/lets-demo/
[36] https://spacy.io
[37] https://stanfordnlp.github.io/stanza/
[38] https://ufal.mff.cuni.cz/udpipe
[39] https://opennlp.apache.org
[40] http://lindat.mff.cuni.cz/services/udpipe/run.php
[41] https://weblicht.sfs.uni-tuebingen.de/weblicht/
[42] https://translate.google.com
[43] https://www.bing.com/translator/
[44] https://www.deepl.com/nl/translator
[45] https://webgate.ec.europa.eu/etranslation/public/welcome.html
[46] https://huggingface.co/yhavinga/gpt2-large-dutch

D'Halleweyn, 2013), an overarching initiative coordinated by the Dutch Language Union to provide the essentials for Dutch language technology, which ran from 2004 till 2011.

The Nederlandse AI Coalitie (Dutch AI Coalition) lists many different use cases, amongst which *Nederlandse AI voor het Nederlands* (Dutch AI for Dutch).[47] The aim of the project is to make speech technology available to everyone who speaks Dutch and not to be dependent on the arbitrariness of large foreign commercial parties. The ambition is to join forces and, as the Netherlands itself, make major improvements in speech technology, especially because collecting and transcribing relevant training material is not feasible for every individual Dutch organisation. The Nederlandstalige Spraak Coalitie (Dutch Speech Coalition)[48] is an initiative to develop speech technology in the Dutch language area, together with various organisations, companies, universities and institutions, as a public-private partnership The Nederlandse Organisatie voor Taal- en Spraaktechnologie – NOTaS (Dutch Organisation for Language and Speech Technology)[49] enables the various players in the field (research institutes, business and government) to join forces to ensure that the Dutch and Dutch-speaking LT industry do not end up in oblivion, but rather lead the way in technological developments.

Computational Linguistics in the Netherlands is a yearly conference that aims to be the meeting point for language technology researchers in the Netherlands and Flanders. The Computational Linguistics in the Netherlands Journal (CLIN Journal)[50] is linked to this conference and provides an international forum for the open access publication of high-quality scholarly articles in all areas of computational linguistics, language and speech technology with special attention for research related to the Dutch language. All published papers are open access and freely available online.

Belgium NLP Meetup[51] is a Belgium-based group for anyone interested in Natural Language Processing. In the meetups, they give a stage to researchers and industry experts that apply NLP in industry and/or academia. They invite everyone with an interest in NLP and related domains (text mining, artificial intelligence, data science, etc.) to join. They are not necessarily focused on Dutch.

The *Common Language Research Infrastructure* (CLARIN) is a European research infrastructure in which the Netherlands play an important role.[52] Flanders has recently (2021) joined CLARIN again through the newly founded CLARIN-BE (Belgium).[53] The Dutch CLARIN Portal Pages CLAPOP[54] bring together all relevant resources created in CLARIN NL and CLARIAH NL projects. The CLARIN portal page at INT[55] provides access to CLARIN tools from the Netherlands and Flanders as INT is a CLARIN technical centre for both. In both the Netherlands and Flanders CLARIN is part of the larger CLARIAH initiative, in which CLARIN and DARIAH, an infrastructure for arts and humanities join forces.

The Flanders AI programme[56] has a research track devoted to NLP, especially Conversational Agents for Dutch in which most universities cooperate.

There are about 15 universities and 70 other organisations providing language technology or resources in the Netherlands. In Belgium there are about 7 universities and 50 other organisations providing LT tools or resources. Apart from the international infrastructures in Section 4 an important source for Dutch language materials can be found at Taalmaterialen[57] from INT. This catalog contains resources, data and tools for linguistic research and language

[47] https://nlaic.com/use-cases/nain-nederlandse-ai-voor-het-nederlands/
[48] https://www.spraakcoalitie.nl
[49] https://notas.nl
[50] https://www.clinjournal.org
[51] https://www.meetup.com/nl-NL/Belgium-NLP-Meetup/
[52] https://www.clarin.eu
[53] https://clarin-be.ivdnt.org
[54] https://portal.clarin.nl/CLAPOP
[55] https://portal.clarin.inl.nl
[56] https://www.flandersairesearch.be/en
[57] https://taalmaterialen.ivdnt.org

and speech technology within the Dutch language area. The Language Machines website[58] of Radboud University contains a plethora of different language technology webservices.

5 Cross-Language Comparison

The LT field[59] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[60] broadly categorised into a number of core LT application areas:
  - Text processing (e. g., part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g., search and information mining)
  - Translation technologies (e. g., machine translation, computer-aided translation)
  - Natural language generation (e. g., text summarisation, simplification)
  - Speech processing (e. g., speech synthesis, speech recognition)
  - Image/video processing (e. g., facial expression recognition)
  - Human-computer interaction (e. g., tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

[58] https://webservices.cls.ru.nl
[59] This section has been provided by the editors.
[60] Tools tagged as “language independent” without mentioning any specific language are not taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

5.2 Levels of Technology Support

We measured the relative technology support for 87 national, regional and minority European languages with regard to each of the dimensions mentioned above based on their respective coverage in the ELG catalogue. For the types of resources and application areas, the respective percentage of resources that support a specific language over the total number of resources of the same type was calculated, as well as their average. Subsequently each language was assigned to one band per resource type and per application area and to an overall band, on a four-point scale, inspired by the scale used in the META-NET White Paper Series, as follows:

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[61]

The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.

5.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. The ELG platform harvests several major LR/LT repositories[62] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[63]

For the purposes of high-level comparison in this report, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific

[61] The thresholds for defining the four bands were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters (i. e., 3%, 10% and 30%) were then used to define the bands per application area and resource type.

[62] At the time of writing, ELG harvests ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and HuggingFace.

[63] Interactive comparison visualisations of the technology support of Europe's languages will be possible on the ELG website using a dedicated dashboard, which dynamically analyses the resources available in the ELG repository, from the middle of 2022 onwards.

level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

That said, we consider the current status of the ELG repository and the higher level findings below adequately representative with regard to the current existence of LT resources for Europe's languages.

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 1 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[64] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

The results of the present comparative evaluation reflect, in terms of distribution and imbalance, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). The complexities of the analyses clearly differ across 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e.g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has

[64] In addition to the languages listed in Table 1, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

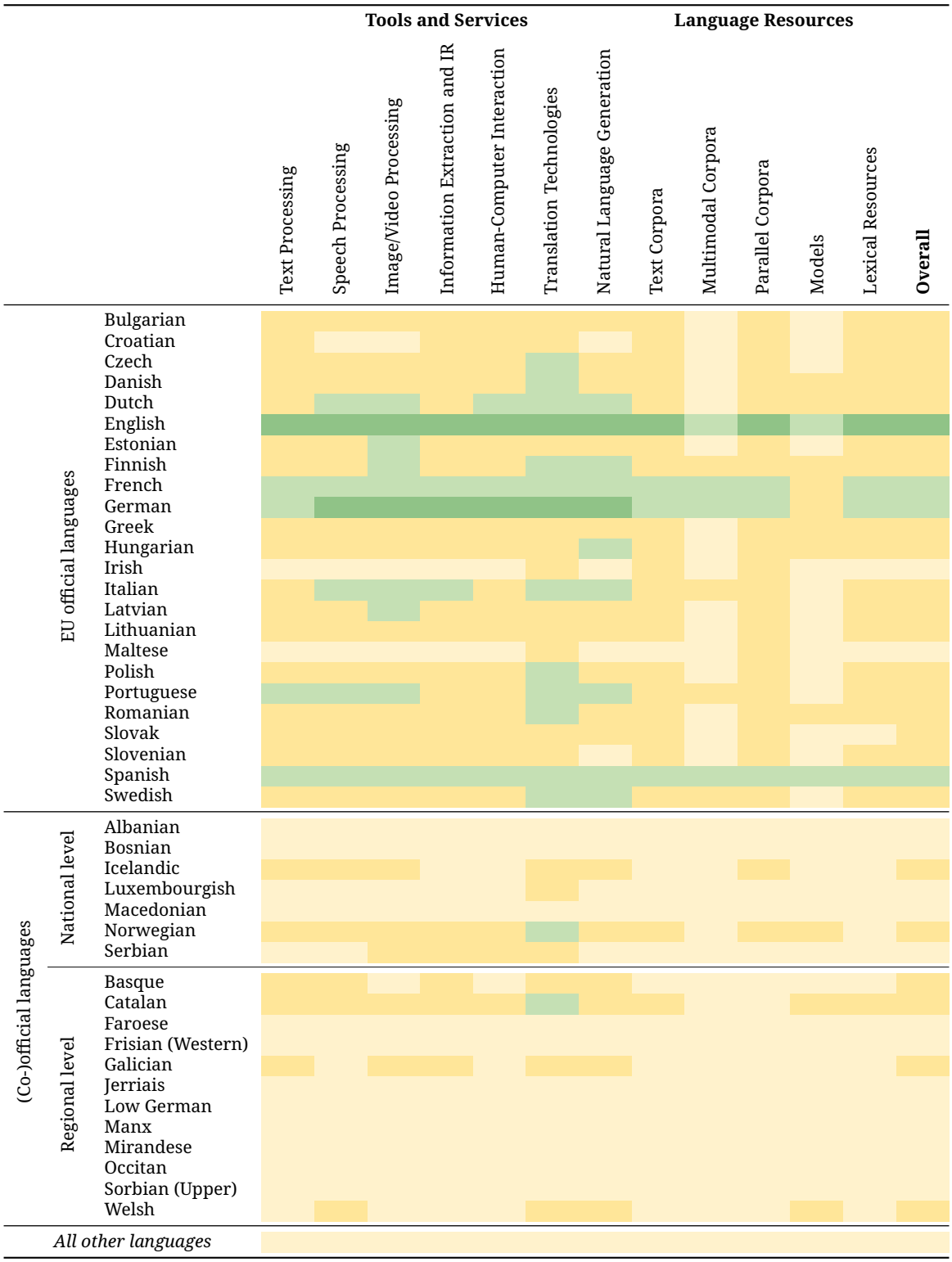

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

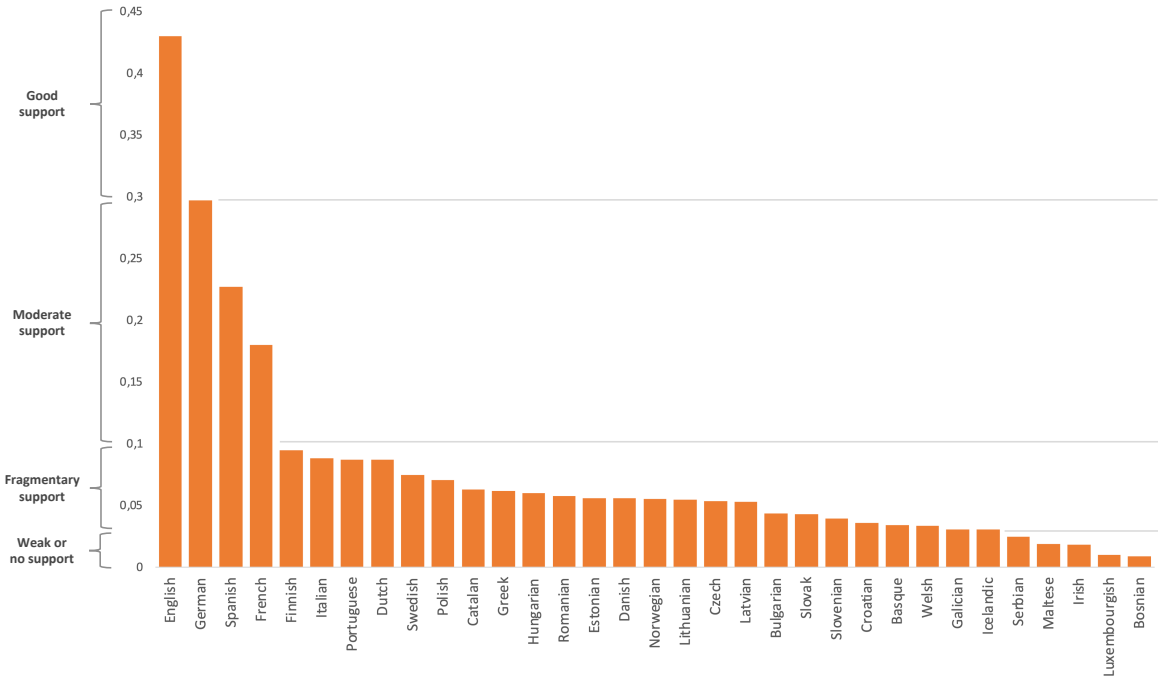

Figure 1: Overall state of technology support for selected European languages (2022)

significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally supported ones is still evidenced in 2022. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality and avert the risks of digital extinction.

6 Summary and Conclusions

Dutch, as the largest of the small Germanic languages, is not in a bad shape digitally. Plenty of data sets and tools are available, and the uptake of Dutch as a language in major NLP applications seems ensured, as witnessed by the inclusion of Dutch in the major online translation engines.

Many of the publicly available open tools rely on publicly available open data sets, and many of these data sets have been created in the STEVIN programme, which lasted till 2011. This implies that these tools have not been adapted to work with the language as used in the last decade. As language use can change rather quickly when new domains become salient in society, it is important to track these changes and allow the tools to learn from recent language use. Therefore, it is paramount that a new programme is set up in which researchers from the Netherlands and Flanders cooperate in the design and construction of corpora that document recent language, be it in written, spoken, or microblog form.

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. European Language Equality D1.2: Report on the state of the art in Language Technology and Language-centric AI, September 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf.

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. European Language Equality D3.1: Report on existing strategic documents and projects in LT/AI, November 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

Noam Chomsky. *Syntactic structures*. The Hague: Mouton, 1957.

CLARIN-D/SfS-Uni. Tübingen. WebLicht: Web-Based Linguistic Chaining Tool. Online, 2012. Date Accessed: 1 Dec 2021. URL https://weblicht.sfs.uni-tuebingen.de/ .

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Stefanie Marzo. Limburgse sjtijl. over het ontstaan en de verspreiding van citétaal. In G. De Sutter, editor, *De vele gezichten van het Nederlands in Vlaanderen. Een inleiding tot de variatietaalkunde.*, pages 291–309. Acco, Leuven, 2017.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Nelleke Oostdijk, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat, and R. Harald Baayen. Experiences from the spoken dutch corpus project. In *LREC*. European Language Resources Association, 2002.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. *The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch*, pages 219–247. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-30910-6. doi: 10.1007/978-3-642-30910-6_13. URL https://doi.org/10.1007/978-3-642-30910-6_13.

Hans Paulussen, Lieve Macken, Willy Vandeweghe, and Piet Desmet. *Dutch Parallel Corpus: A Balanced Parallel Corpus for Dutch-English and Dutch-French*, pages 185–199. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-30910-6. doi: 10.1007/978-3-642-30910-6_11. URL https://doi.org/10.1007/978-3-642-30910-6_11.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. Open Dutch WordNet. In *Proceedings of the Eight Global Wordnet Conference*, Bucharest, Romania, 2016.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Tilly Ruitenberg, Jesse De Does, and Katrien Depuydt. Developing gigant, a lexical infrastructure covering 16 centuries. In Anne Dykstra and Tanneke Schoonheim, editors, *Proceedings of the 14th EURALEX International Congress*, pages 468–476, Leeuwarden/Ljouwert, The Netherlands, jul 2010. Fryske Akademy. ISBN 978-90-6273-850-3.

Peter Spyns and Elisabeth D'Halleweyn. The STEVIN Programme: Result of 5 Years Cross-border HLT for Dutch Policy Preparation. In Spyns P. and Odijk J., editors, *Essential Speech and Language Technology for Dutch.*, pages 21–39. Springer, Berlin, Heidelberg, 2013.

Frieda Steurs. Nederlands een grote taal? Een kewstie van meten. *Neerlandica Wratislaviensia*, pages 17–29, 2021.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Stéphan Tulkens, Chris Emmery, and Walter Daelemans. Evaluating unsupervised dutch word embeddings as a linguistic resource. In *LREC*. European Language Resources Association (ELRA), 2016.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

A. van den Bosch, G.J. Busser, S.V.M. Canisius, and W. Daelemans. *An efficient memory-based morphosyntactic tagger and parser for Dutch*, pages 191–206. LOT, 2007. Pagination: 16.

Gertjan van Noord. At last parsing is now operational. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*, pages 20–42, Leuven, Belgique, April 2006. ATALA. URL https://aclanthology.org/2006.jeptalnrecital-invite.2.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. *Large Scale Syntactic Annotation of Written Dutch: Lassy*, pages 147–164. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-30910-6. doi: 10.1007/978-3-642-30910-6_9. URL https://doi.org/10.1007/978-3-642-30910-6_9.