TITLE
The CELEX Lexical Database: History, Structure, Applications, and Accessibility

AUTHOR
Marc van Oostendorp
Modern Languages and Cultures
Radboud University
PO Box 9103
6500 HD NIJMEGEN
The Netherlands

Abstract
The CELEX Lexical Database is a comprehensive multilingual resource developed in the Netherlands during the late 1980s, providing richly annotated lexicons for English, Dutch, and German. This paper outlines its development history, database structure, and applications in psycholinguistics, computational linguistics, and speech technology. Despite newer resources emerging, CELEX remains valuable for researchers requiring detailed linguistic annotations including orthography, phonology, morphology, syntax, and frequency information. Maintained by the Max Planck Institute for Psycholinguistics, it continues to bridge theoretical, psychological, and computational approaches to language research.

Keywords
Computational linguistics, corpus linguistics, lexical database, morphology, phonology, psycholinguistics, speech technology, word frequency

Keypoints

- Trace the historical development of CELEX from its Dutch origins to international adoption
- Describe the comprehensive structure of the trilingual database (English, Dutch, German)
- Detail the linguistic information provided for each lexical entry (orthography, phonology, morphology, syntax, frequency)
- Highlight CELEX's applications in psycholinguistics, computational linguistics, and speech technology
- Analyze CELEX's current status and continued relevance in the context of newer lexical resources
- Assess CELEX's legacy and contribution to language research methodology

Draft version. Full version appeared as van Oostendorp, M. (2026). The CELEX Lexical Database: History, structure, applications, and accessibility. In H. Nesi & P. Milin (Eds.), *International encyclopedia of language and linguistics* (3rd ed., pp. 386–390). Elsevier. https://doi.org/10.1016/B978-0-323-95504-1.01259-X

1. Introduction

The CELEX Lexical Database is a comprehensive multilingual lexical resource originally developed in the Netherlands and now hosted at the Max Planck Institute for Psycholinguistics in Nijmegen. It provides richly annotated lexicons for English, Dutch, and German, including detailed information on word forms, their structures, and usage frequencies. Since its creation in the late 1980s and early 1990s, CELEX has been an important tool for linguists, psycholinguists, and language technologists. This article offers an overview of CELEX's history, the structure and contents of its databases, its applications in various fields, and how researchers can access it. Comparisons with other well-known linguistic resources (such as WordNet, CHILDES, and SUBTLEX) are also included to contextualize CELEX's role in the landscape of language data.

2. History of CELEX

The CELEX project (an acronym for Centre for Lexical Information) began in the mid-1980s as a national initiative to create an electronic lexical database for research. Funding was provided by the Dutch government (e.g. the Netherlands Organisation for Scientific Research, NWO, and the Ministry of Science and Education), recognizing the need for accessible, machine-readable lexical data. The development was a joint enterprise involving multiple Dutch institutions: the University of Nijmegen, the Institute for Dutch Lexicology (INL) in Leiden, the Max Planck Institute for Psycholinguistics (MPI) in Nijmegen, and the Institute for Perception Research (IPO) in Eindhoven.
Work on CELEX was underway by the late 1980s. A prototype of the Dutch lexical database was accessible by 1987, followed by an English database in 1988, and the system became fully operational around 1989. The project's first full release on CD-ROM came in the early 1990s. R. Harald Baayen, Richard Piepenbrock, and Huib van Rijn compiled the first CD-ROM edition of CELEX (published in 1993). A board chaired by prominent psycholinguist Willem Levelt (MPI) oversaw the project, ensuring collaboration between the participating institutes.
The first CELEX release (early 1990s) was followed by a second release in 1995 (often referred to as CELEX2). This CELEX2 update included expansions and corrections, especially to the German database. Additionally, frequency counts were integrated more directly into the data files, and syllable frequency statistics were introduced for Dutch and English.
The Linguistic Data Consortium (LDC) collaborated in publishing the CD-ROMs (handling pre-mastering and distribution). Since the 1995 release, the content of CELEX has remained largely static, reflecting the state of the lexicons and corpora at that time. Nonetheless, CELEX's thorough design has ensured its continued relevance in research well into the 2000s.

2. Structure of the Database

2.1 Languages

CELEX actually comprises three parallel lexical databases – one each for British English, Dutch, and German. Each database is structured in a similar way and contains

extensive information for tens of thousands of words in that language. For reference, the second-release CELEX (circa 1995) contains approximately:

- Dutch: 124,136 lemmas and 381,292 wordforms.
- English: 52,446 lemmas and 160,594 wordforms.
- German: 51,728 lemmas and 365,530 wordforms.

Each “lemma” entry corresponds to a headword (typically a citation form or base form of a word), while “wordform” entries cover inflected or derived forms. CELEX distinguishes lemma lexicons and wordform lexicons for each language, linking them via unique identifiers. Users of the database can choose to work at the lemma level or the wordform level; a separate “corpus lexicon” is also provided, which lists all word tokens (types) observed in the source corpora with their frequencies. This separation allows flexible queries – one can retrieve information about base forms, examine all inflected forms of a lexeme, or explore raw corpus frequencies for every distinct word form.

A major strength of CELEX is the rich set of linguistic attributes it provides for each entry. The database was designed to be “multifunctional and polytheoretical,” containing multiple layers of linguistic description. Key components of the lexical entries include:

Orthography: Standard spellings of each word (with diacritics where relevant), alternative spelling variants, and hyphenation positions (useful for text formatting). CELEX even encodes alphabetically sorted forms (for instance, a version of the word’s spelling with all letters in lowercase and diacritics removed for consistent sorting).

Phonology: Pronunciations in phonetic transcription. For each word, CELEX provides one or more phonemic transcriptions (in a machine-readable phonetic alphabet) along with syllable boundaries and primary stress markers. Variants in pronunciation (e.g. dialectal or alternate pronunciations) are included when available. Phonological information also covers syllable structure details (such as consonant/vowel patterns) and phoneme frequency data.

Morphology: Detailed morphological analysis is given. Each complex word is broken down into its constituent morphemes (roots and affixes) or into compound components. CELEX indicates derivational relationships (how words are formed via prefixes, suffixes, etc.), compound structures (for languages like Dutch and German which have many compound words), and full inflectional paradigms for each lemma. For example, a verb entry will list all conjugated forms (wordforms) along with a code indicating its inflection class; a noun entry will show plural and possessive forms, etc. This morphological parsing was done largely by hand by linguists, ensuring a high-quality decomposition of word structure.

Syntax: Basic syntactic information is recorded for each lemma. This includes the part of speech (word class) and subcategorization information. For verbs, CELEX provides argument structure codes indicating what types of complements the verb can take (for instance, whether a verb is transitive, intransitive, ditransitive, etc., and what prepositions might be required). Nouns and adjectives include codes for relevant grammatical features or complementation patterns. These syntactic annotations allow users to, for example, extract all verbs that take a certain argument structure, or all nouns of a certain gender (in German) and so on.

Word Frequency: Each entry (lemma and wordform) is accompanied by frequency counts derived from large text corpora. CELEX provides both lemma frequency (how common the headword is, summing all its forms) and wordform frequency (how common that specific form is). These counts are based on "recent and representative text corpora" of the respective languages. Frequency data is typically given as raw counts and/or normalized frequencies (e.g. per million words). This makes CELEX not just a dictionary, but also a source of usage statistics, which is invaluable for psycholinguistic research.

2.2. Data Sources and Compilation

The contents of the CELEX lexicons were compiled from various authoritative sources, ensuring broad coverage of each language's vocabulary.

For Dutch, the compilers drew from major dictionaries and a text corpus: Van Dale's Comprehensive Dictionary of Contemporary Dutch (1984) contributed ~80k lemmas, the official Dutch word list ("Groene Boekje", 1954) added ~65k lemmas, and an additional ~15k high-frequency words came from the INL corpus (a 42 million word corpus of Dutch). Due to overlaps between sources, the final Dutch lemma list was about 124k, covering an estimated 95% of word tokens in the INL corpus.

For English, CELEX incorporated lemmas from two well-known learners' dictionaries: the Oxford Advanced Learner's Dictionary (1974) (~41k entries) and the Longman Dictionary of Contemporary English (1978) (~53k entries). After accounting for overlaps (around 30k words appear in both sources), the English lexicon ended up with ~52k lemmas. No additional text corpus lemmas were added for English, but the coverage was still about 92% when compared to a ~18 million word corpus (the Birmingham/COBUILD corpus). The English word frequencies in CELEX are derived from that COBUILD corpus (which included British and some American texts from the 1980s).

For German, the approach was slightly different: rather than published dictionaries, CELEX used digital lexical lists such as Bonnlex and Molex (machine-readable word lists from the University of Bonn and the Institut für Deutsche Sprache in Mannheim) and a German spelling lexicon from MIT's Noetic Circle Services. These sources collectively provided stems, inflected forms, and lemmas for German. The German lemma count (~51k) was smaller, covering roughly 83% of a 6-million-word reference corpus. German frequency counts came from a combination of written texts (5.4 million tokens from newspapers, fiction, non-fiction) and spoken transcripts (0.6 million tokens) available in the early 1990s. Notably, the German texts were drawn from several corpora (e.g., Mannheimer Korpus and Bonner Zeitungskorpus for written language, and the Freiburger Korpus for spoken language).

2.3. Data Format and Organization

Technically, the CELEX database was originally implemented in an Oracle relational database and accessible through a custom interface called FLEX on UNIX systems.

However, the distributed version (on CD-ROM and now via download) is organized as a collection of plain text ASCII files for maximum compatibility. The lexicon is split into multiple files by information type (e.g., separate files for orthography, phonology, morphology, etc.) and by lexical class (lemmas vs wordforms vs corpus list). Each entry is identified by a unique number so that data from different files can be joined together as needed. For example, one can use the unique ID to link a word's orthographic entry to its phonological entry, morphological parse, and frequency count across different files. The CELEX distribution includes documentation and utilities to assist users, such as example AWK scripts for extracting or computing particular fields (e.g., generating sorted spelling forms, or joining data from lemma and wordform tables)article. This design allows researchers considerable flexibility, albeit with some effort to write queries or scripts. The database design is consistent across the three languages, which enables comparative studies if needed (though each language's data is contained in its own directory). The thorough linguistic documentation (CELEX User Guide) defines all codes and fields so users can interpret the data correctly. Overall, the structure of CELEX reflects a careful balance between human-readable organization (through labelled columns and text files) and machine-readable detail (through systematic coding and IDs), making it a durable resource for diverse linguistic analyses.

3. Applications of CELEX

Since its release, CELEX has been utilized in a wide range of research areas and applications in language science and technology. Some of the key domains include:
Psycholinguistics and Cognitive Science: CELEX has been widely used in psycholinguistic research as a source of lexical statistics (e.g., word frequency, word length, neighborhood density, morphological complexity) when designing experiments or analyzing language behavior. For example, reaction time studies of word recognition often control for or examine frequency effects using CELEX frequency counts. The English Lexicon Project and similar large-scale studies include CELEX frequency measures as predictors of human response times. Because CELEX provides lemma frequencies and wordform frequencies, researchers can investigate how whole-word frequency versus base-form frequency influence cognitive processing. CELEX's morphological and phonological data have also been used to study how complex words (like inflected or compound words) are processed by the brain – for instance, examining the role of morphological family size or syllable frequency in word recognition. In short, CELEX has served as an important "normative" database in psycholinguistics, much like how norms for imageability or age-of-acquisition are used, providing baseline lexical properties for thousands of words.
Computational Linguistics and NLP: In computational linguistics, CELEX has been valued for its extensive lexical annotations which can train or evaluate language models. The database's detailed morphological parses have been used in developing and testing morphological analyzers and generators. For example, algorithms for automatic morphological segmentation have been evaluated against CELEX's human-annotated morpheme breakdowns for English, Dutch, or German words. The syntactic subcategorization codes have been used in syntactic parsing research to ensure that parsing models respect the valency of verbs (e.g., knowing which verbs are transitive or not). CELEX can also function as a pronunciation lexicon for NLP tasks: the phonetic

transcriptions (especially for English and Dutch) have been used to create pronunciation dictionaries for speech recognition and text-to-speech systems. Because CELEX includes syllabification and stress information, it has even been used to develop and evaluate algorithms that determine syllable boundaries or stress patterns in text. In the 1990s and 2000s, CELEX was one of the few readily available machine-readable lexicons, so it was often used in corpus linguistics and NLP projects as a reference lexicon (for spell-checking, lemmatization, and other text processing needs).
Speech Technology (ASR/TTS): CELEX has found applications in speech processing, including both automatic speech recognition (ASR) and text-to-speech (TTS) synthesis. The phonological data (phonemic forms with stress) serves as a high-quality pronunciation dictionary. For instance, speech synthesis systems can use CELEX entries to know how to pronounce words not explicitly in their lexicon, and ASR language models can use CELEX to expand vocabulary with probable pronunciations. CELEX's inclusion of multiple pronunciation variants (e.g., British vs American English alternatives, or multiple Dutch variants) and its consistent phonemic encoding make it useful for training pronunciation models. Moreover, frequency information from CELEX helps speech systems decide which words are more likely and thus should perhaps have multiple pronunciations considered. Early research in speech synthesis and speech recognition often leveraged CELEX as a source for phonetic and syllabic rules. Even beyond words, the syllable frequency data included in CELEX (for English and Dutch) has been used to study phonotactics and to generate likely nonwords for speech experiments.
Language Education and Lexicography: Although primarily a research tool, CELEX has occasionally been used in language teaching and lexicography. The detailed word frequency lists in CELEX (especially when it was first released) provided educators and textbook authors with information on which words are most common, helping to select vocabulary for language learning curricula. Lexicographers have used CELEX to verify dictionary entries or to obtain frequency-based evidence for usage notes. Because CELEX covers multiple languages in a unified framework, it has also supported contrastive linguistics studies and the development of bilingual lexical aids (e.g., comparing the morphological complexity of English vs Dutch words).
In summary, CELEX is a highly versatile resource. Its creators envisioned it as serving "theoretical linguistics, psycholinguistics, machine translation, and natural language interfaces", and indeed it has been employed in all those areas. The database's influence is evident from how frequently it is cited in research papers throughout the 1990s and 2000s – for example, as "the CELEX lexical database (Baayen, Piepenbrock & van Rijn, 1993)" in studies of word frequency effects. Its balanced combination of linguistic theory-agnostic data (hence "polytheoretical") with practical corpus-based statistics made it a foundational tool before the era of massive web-derived corpora. Even today, researchers sometimes turn to CELEX for well-curated lexical information that newer resources might lack.

4. Current Status

4.1. Data Formats

The CELEX data are provided online as plain text files (often with the extension .cd for "Celex Database") that can be opened in any text editor or processed with scripts.

Accompanying documentation explains the directory structure and file naming conventions. For instance, there are separate directories for English, Dutch, and German, each containing subfolders for lemmas, wordforms, etc., and within those, files like eng_lemma.txt, eng_wordform.txt (with columns separated by either spaces or specified delimiters). Because the data is not in a single database file, users typically load it into a database or write scripts to query it. The original CELEX user interface (FLEX) is not distributed, so users interact with the data via their own tools.

4.2 User Community

The primary users of CELEX are researchers and scholars in fields like linguistics, psycholinguistics, computational linguistics, and speech and language technology. University departments often maintain a local copy of CELEX for research and teaching. Students in linguistics may use CELEX data for projects (for example, to correlate word length and frequency, or to illustrate morphological phenomena). Some natural language processing companies have also licensed CELEX in the past, particularly for speech or text processing in Dutch and German where comprehensive lexicons were harder to come by. Over time, as newer resources (and larger corpora) have emerged, CELEX's user base has become more specialized — it is not a crowdsourced or continuously updated resource, so general-purpose use (e.g., by casual end-users) is limited. However, for those who need high-quality lexical metadata (like stress-marked pronunciations or handcrafted morphological parses), CELEX remains a go-to reference.

4.3. Maintaining Relevance

It should be noted that CELEX, being a product of the 1980s-90s, reflects the state of language technology and language use up to that period. The frequency counts come from corpora that are now several decades old, and no newer words (neologisms after the 1990s) are included. Researchers must keep this in mind; for example, very modern terms will not appear in CELEX. That said, many core vocabulary items of English, Dutch, and German are stable, and the detailed linguistic information in CELEX is still valid. The resource is often cited with its original reference (Baayen, Piepenbrock & Gulikers or van Rijn, 1993/1995) in contemporary studies, indicating its lasting significance. In recent years, some projects have created derived databases or formats (for instance, converting CELEX into SQL databases or JSON formats, available on platforms like GitHub for those with access) to make it easier to query. These efforts extend the usability of CELEX with modern tools, but they all trace back to the official data.

5. Conclusion

In conclusion, the CELEX lexical database stands as a landmark resource in the history of computational linguistics and psycholinguistics. Historically, it was pioneering: a project that brought together experts and institutions to create a reusable lexical knowledge base well before the age of "big data." Its design reflects careful thinking about what information about words is most useful – covering everything from how

words are spelled and pronounced to how they are structured and used in context. Structurally, CELEX's comprehensive and standardized format for three major languages made it a powerful tool for cross-linguistic lexical studies and applications. Practically, it has fueled research and development in many areas: helping scientists understand human language processing and providing engineers with data for language technologies. CELEX's influence is also seen in how it set a precedent for later resources; for instance, many modern lexicons and corpora have expanded upon the groundwork that CELEX laid (sometimes surpassing it in size or specialization, as SUBTLEX did for frequency data, for example). As of 2025, CELEX remains available through the MPI and LDC for those who need its depth of lexical detail. While it may no longer be the go-to source for the very latest vocabulary or the largest corpora, it retains a core role in linguistic research – especially for studies requiring high-quality, linguist-annotated lexical information that raw "big data" corpora might not easily provide. In an era when resources like WordNet, CHILDES, and SUBTLEX each address specific needs (semantic networks, child language data, updated frequencies), CELEX continues to occupy a complementary niche by offering a well-curated, multi-faceted lexicon. Its enduring usage in academic papers and its continued hosting at a major institute testify to its value. In sum, the CELEX database is not just a relic of 20th-century linguistics, but a lasting reference point for anyone interested in the building blocks of English, Dutch, and German words – a true legacy resource bridging theoretical, psychological, and computational approaches to language.

References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.
Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX2 [CD-ROM]. Linguistic Data Consortium*, Philadelphia.
CELEX Readme and User Guides (1995)
– Documentation accompanying the CELEX CD-ROM, detailing the database content, sources, and structure.
LDC Catalog Entry for CELEX2 (1996)article
– Official description of CELEX's contents and applications.
LINGUIST List announcement (1997)linguistlist.org
– Information on the release of CELEX second edition, including size and access.
SUBTLEX documentation (Brysbaert & New, 2009)
ugent.be
– Discussion of CELEX vs other frequency norms in predicting word recognition performance.
en.wikipedia.org
– Contextual information on other linguistic resources for comparison.
StackExchange post (2014)
english.stackexchange.com
– Note on the MPI online interface for querying CELEX.