OXFORD
UNIVERSITY PRESS

Oxford Languages

↓

About us

OXFORD
UNIVERSITY PRESS

Oxford University Press is the world's largest university press. Today OUP has offices in 50 countries.

OxfordLanguages

Looking beyond traditional publishing to develop new ways of supporting customers, OUP established Oxford Languages to provide digital languages data.

Oxford Dictionaries

↓

OxfordLanguages

We rebranded to better reflect the changing needs of our customers, moving beyond traditional dictionary publishing into human-curated language data provision.

Language data by language specialists

Data for AI

Our language data specialists build unique lexical (text and speech) datasets suitable for model training and other natural language processing (NLP) applications.

OxfordLanguages

Machine Translation

Parallel datasets that can support machine translation.

AI voice generator

Pronunciation datasets with lexical transcriptions and audio to improve text-to-speech and AI dubbing applications.

Conversational AI

Language databases designed to help with natural language understanding, enabling models to learn languages and interpret meaning accurately.

AI writing assistant tools

Lexical datasets that aid writing tools in suggesting grammar, spelling, and vocabulary improvements.

NLP Keyboard

Sensitivity labels in the data can be used to improve handling of offensive, vulgar, or demeaning language, while dialect labels improve text prediction in regional dialects and language variations.

Use cases we *support*

Our clients

Microsoft Translator

WELLSAID

Google Translate

texthelp®

AI21 labs

synthesia

ELEMENTAL cognition

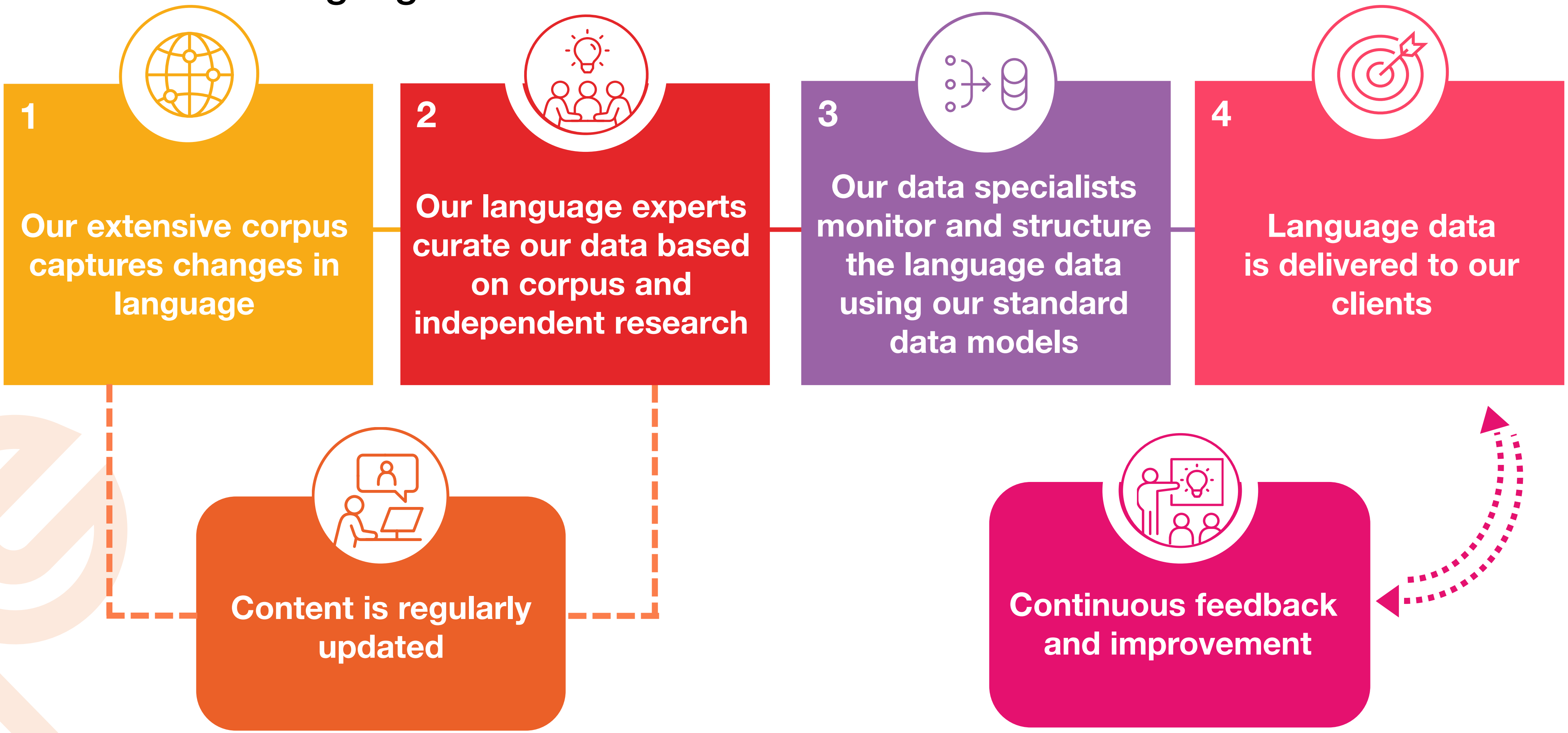
grammarly

Available Languages

- Afrikaans
- Arabic
- Assamese
- Bengali
- Bulgarian
- Catalan
- Chinese Simplified
- Chinese Traditional
- Croatian
- Czech
- Danish
- Dutch
- English American
- English Australian
- English British
- English Canadian
- English Indian
- Finnish
- French Canadian
- French European
- Georgian
- German
- German Swiss
- Greek Modern
- Gujarati
- Hausa
- Hebrew Modern
- Hindi
- Hungarian
- Indonesian
- isiXhosa
- isiZulu
- Italian
- Japanese
- Kannada
- Kazakh
- Korean
- Latvian
- Malay
- Malayalam
- Marathi
- Northern Sotho
- Norwegian
- Odia
- Polish
- Portuguese Brazilian
- Portuguese European
- Punjabi
- Quechua
- Romanian
- Russian
- Setswana
- Slovenian
- Spanish European
- Spanish Latin American
- Swahili
- Swedish
- Tamil
- Tatar
- Telugu
- Tok Pisin
- Thai
- Turkish
- Turkmen
- Ukrainian
- Urdu
- Vietnamese
- Welsh

Accurate and reliable data you can trust

Our datasets' language content is carefully curated and annotated by language experts who are passionate about language.



1 Our extensive corpus captures changes in language

2 Our language experts curate our data based on corpus and independent research

3 Our data specialists monitor and structure the language data using our standard data models

4 Language data is delivered to our clients

Content is regularly updated

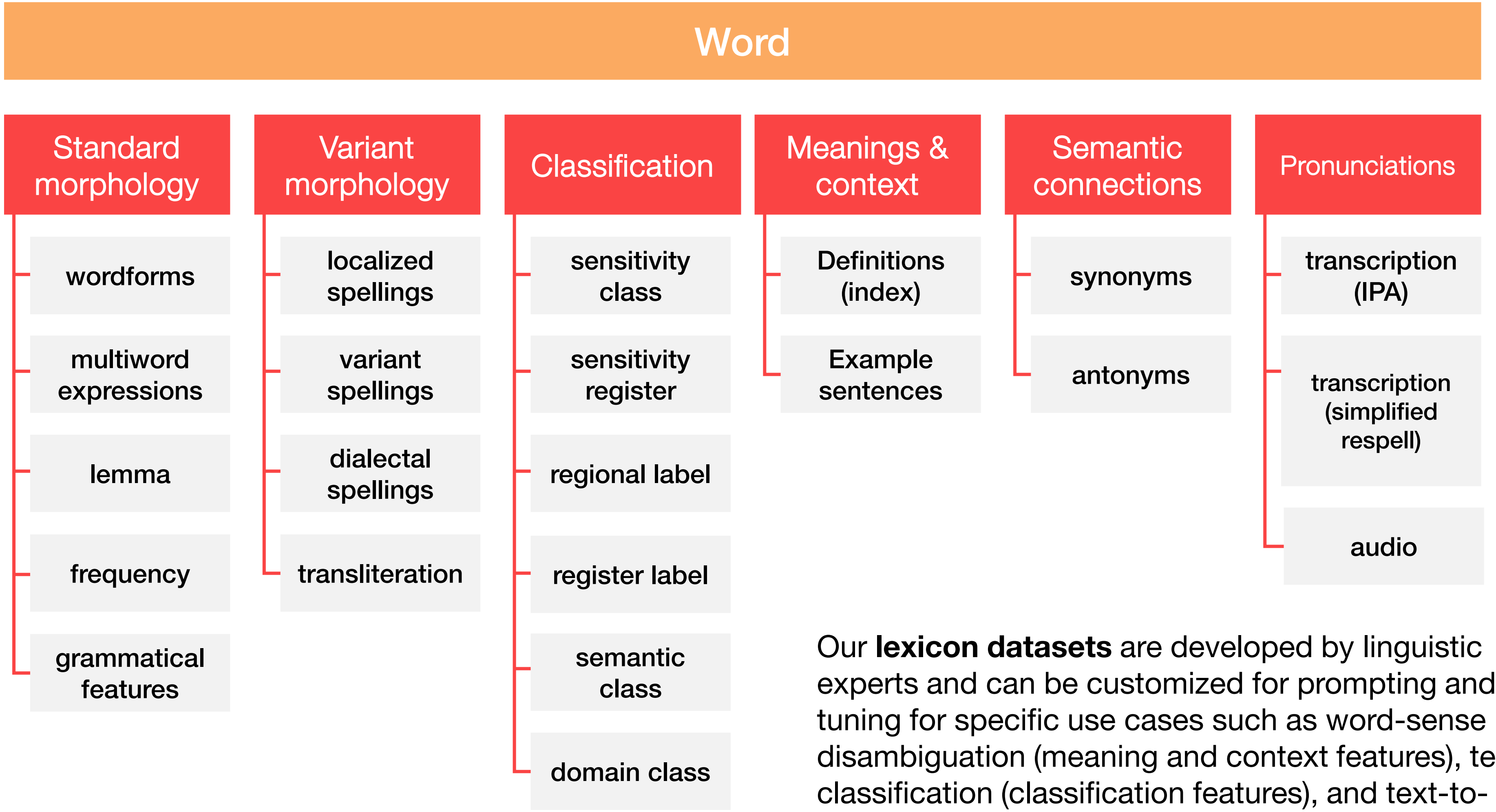
Continuous feedback and improvement

Lexical
Datasets

Built using a combination of world-leading corpus data, human-curated dictionary content, and reviewed by native linguists at every step, our lexical data is optimized for NLP solutions.

OxfordLanguages

Lexical datasets features



Word

- Standard morphology
  - wordforms
  - multiword expressions
  - lemma
  - frequency
  - grammatical features
- Variant morphology
  - localized spellings
  - variant spellings
  - dialectal spellings
  - transliteration
- Classification
  - sensitivity class
  - sensitivity register
  - regional label
  - register label
  - semantic class
  - domain class
- Meanings & context
  - Definitions (index)
  - Example sentences
- Semantic connections
  - synonyms
  - antonyms
- Pronunciations
  - transcription (IPA)
  - transcription (simplified respell)
  - audio

Our **lexicon datasets** are developed by linguistic experts and can be customized for prompting and fine-tuning for specific use cases such as word-sense disambiguation (meaning and context features), text classification (classification features), and text-to-speech (pronunciation features).

Basic features and sample

philologien
(Wordform)

philologie
(Lemma)

Noun
(Part of Speech)

Feminine, plural, dative.
(Grammatical features)

Sample of one lemma 'philologie' from a basic German dataset.

Wordform	Lemma	PoS	Gender	Number	Case
philologie	philologie	N	Fem	Sg	Acc
philologie	philologie	N	Fem	Sg	Dat
philologie	philologie	N	Fem	Sg	Gen
philologie	philologie	N	Fem	Sg	Nom
philologien	philologie	N	Fem	Pl	Acc
philologien	philologie	N	Fem	Pl	Dat
philologien	philologie	N	Fem	Pl	Gen
philologien	philologie	N	Fem	Pl	Nom

Features for specific use cases

Use case specific features	Features description	Example of languages with this type of feature	Useful for
Domain classification with examples	Our lexical datasets present the domain classifications for wordforms, and also domain-relevant examples for words with more than one classification to support disambiguation.	All languages	Text classification
Spelling variants	In many languages, spelling is not standardized, and therefore spelling of the same words can vary. Our lexical datasets present a core wordform (which follows a defined specification), and this is supported by additional spelling variant features which presents the most likely alternative spellings.	Hindi, English	Localization, assisted writing
Frequency per wordform, frequency per locale (e.g., American English, Canadian English)	Our lexical datasets present the normalized frequency of each wordform in specified general corpora, or in a relevant region-specific corpus to support localized user experiences.	All languages incl. English (World English varieties)	Localization, assisted writing
Sensitivity class and register	These features support detection of offensive vulgar, and potentially sensitive (in certain contexts) words within NLP pipelines. Our lexical datasets present wordform-level sensitivity classification and categorization (e.g., drug abuse, body part).	All languages	Text classification, hate speech detection
Dialectal translations	In regions where there are multiple spoken dialects and informal spellings of these dialects, our lexical datasets deliver a solution which show the equivalences between dialects, and to allow you to track back to a 'standard' variety of the language which can be understood by all dialects.	Swiss German	Localization/ de-localization
Transliterations	More tech users want to interchangeably use their native and Roman scripts in their experience. Our lexical datasets present a solution to this by presenting the native script wordform with the equivalent roman wordforms as the 'transliteration' feature.	Hindi, Tamil, other Indian languages, Japanese, Chinese (Simplified and Traditional)	Assisted writing
Pronunciations: IPA (or other transcription system) and audio	Our lexical datasets present accurate, localized written (IPA and respell transcriptions) and audio pronunciations of the wordforms.	All languages incl. World varieties of English and Spanish, Indian languages	Text-to-speech, speech recognition

- Complete coverage of World English varieties:
    - British
    - American
    - Indian
    - Australian
    - Canadian
- Normalized frequency of each wordform in specified general corpora, or in a relevant region-specific corpus.

Australia
Canada
India

English Localization

Increase the range and diversity of models with localized lexical data covering English beyond the UK and US.

Pronunciation data

Wordform example: friendliness (US English)

Audio pronunciation

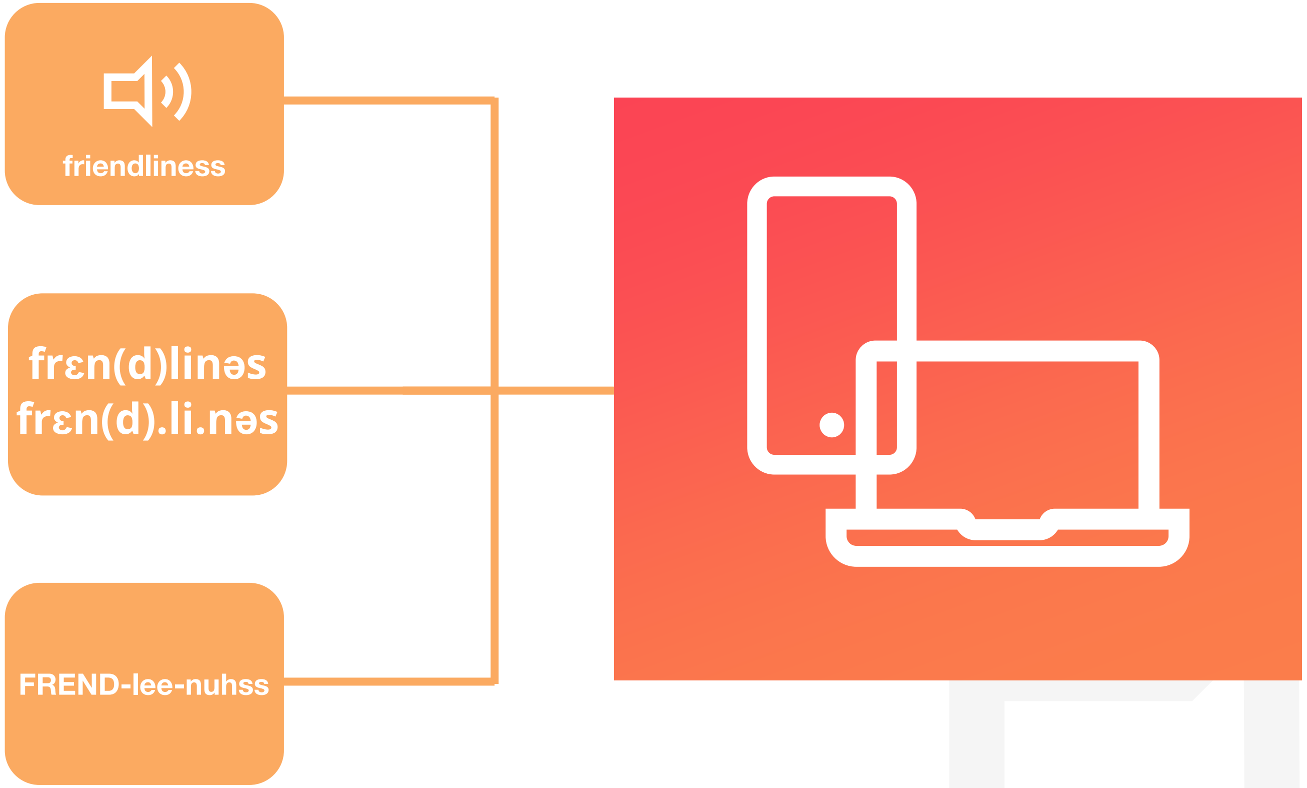
High quality audio files recorded in a controlled environment.

International phonetic alphabet transcription (IPA)

Pronunciation transcription in IPA to Oxford Languages style, with or without syllabification.

Oxford English Simple Text Respell

Provides a visual means for interpreting pronunciations, without the need for phonetic or linguistic knowledge.



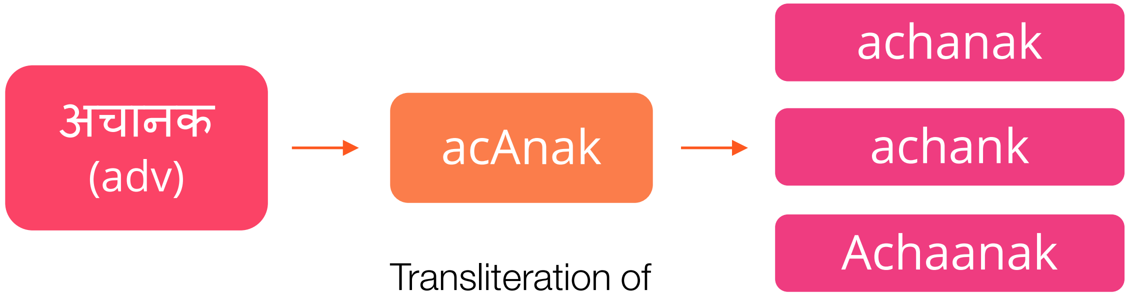
Our pronunciation datasets include coverage of English spoken in other parts of the world, such as **American English, British English, Indian English and Australian English.**

OxfordLanguages

Hindi

Featuring transliterations and spelling variants, our data allows models to process how Hindi is written and spoken today.

- Lexical data that covers the most important words for Hindi speakers.
- Spelling variation specification that represents the breadth of spelling used by Hindi speakers across India.
- Strict and colloquial transliteration.



अचानक (adv) → acAnak → achanak / achank / Achaanak

In Devanagari which means ‘suddenly’

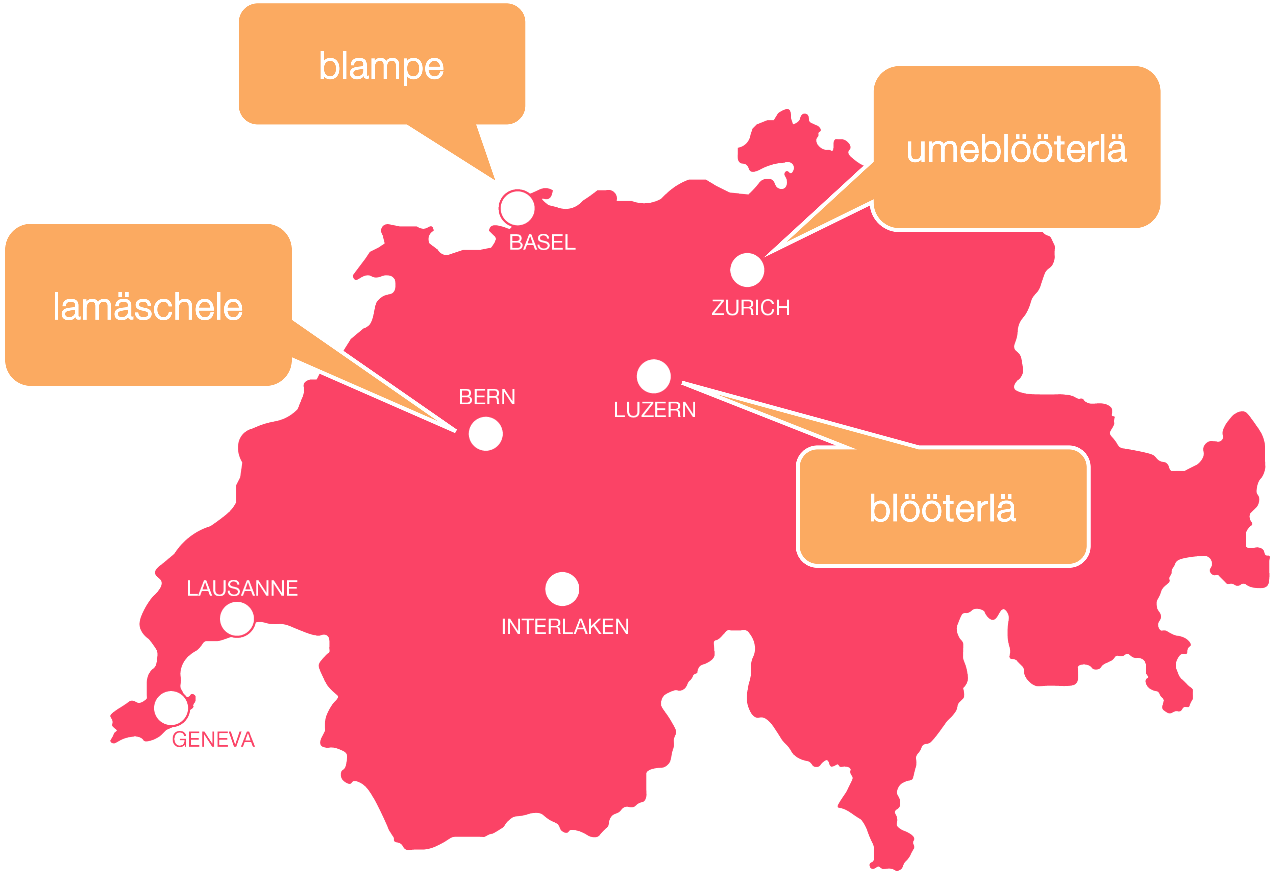
Transliteration of the word which is accurate to the pronunciation.
It can be used for text-to-speech.

Transliterations which are used by Hindi speakers in informal writing such as social media.
It can be used for assistive writing and other NLP use cases.

OxfordLanguages

Swiss Standard German: ***langsame*** (plural of langsam, meaning *slow* or *slowly* in SSG)

Below are Swiss German dialects for the word ***langsame.***



Swiss German

Dialectal varieties data focused on presenting Standard Swiss German and the Swiss German dialects spoken in Bern, Basel, Zurich and Luzern.

Swiss German

Further examples:

Swiss Standard German Wordform	Bern	Basel	Zurich	Luzern
langsamen	lamäschele	blampee	umeblööterläe	blööterlää
langsame	lamäscheli	blampi	umeblööterläi	blööterläi
langsamen	lamäscheleä	blampee	umeblööterläe	blööterlää
langsame	lamäscheli	blampi	umeblööterläi	blööterläi
langsamen	lamäscheleä	blampee	umeblööterläe	blööterlää
langsamen	lamäschelei	blampei	umeblööterläi	blööterläi

Academic Corpus

An archive of over 2 billion words and a continually updated pipeline of published content representing the highest standards in academic publishing and covering a diverse range of research domains and genres.

Archival corpus of academic research books

Available for pretraining of language models

TOTAL # OF POSSIBLE TITLES
19,300+

AVERAGE PAGE COUNT
338

EST. # OF WORDS
3.38 Billion

EST. # OF SENTENCES
169 Million

Language variety
British & American English

Arts & Humanities
6,755 titles

Social Sciences
6,755 titles

Medicine and Health
2,123 titles

Science and Mathematics
1,351 titles

Law
2,316 titles

Numbers are approximate as of July 2023

Parallel Sentences

We offer English sentences that are optimized for translation. These sentences are translated into multiple languages and can be used as training and validation datasets for machine translation.

OxfordLanguages

Parallel Sentences

English source sentences cover a variety of simple, complex, and colloquial sentences, with length of sentences ranging from 4-25 words

Translation is completed by native speaker linguists to a defined translation specification, which favors natural translation.

The persistent semantic inventory identifier (OUPLexID) which is associated with a specific word-sense allows engineers to utilize other sense-specific lexical data from other data in the Oxford Languages linked data ecosystem.

Oxford Sentence Dictionary

Our largest sense-annotated dataset of real-life examples of English in use.

Contains over 1.9 million sentences representing over 200,000 distinct meanings of over 90,000 words.

It provides up to 20 examples for each meaning, giving a broad range of examples.

OxfordLanguages

Thank you

Any *questions?*

✉ Strategic Account Manager
scott.hamilton@oup.com

✉ Strategic Account Manager
alistair.butchers@oup.com

Oxford Languages website

OXFORD
UNIVERSITY PRESS

Oxford Languages