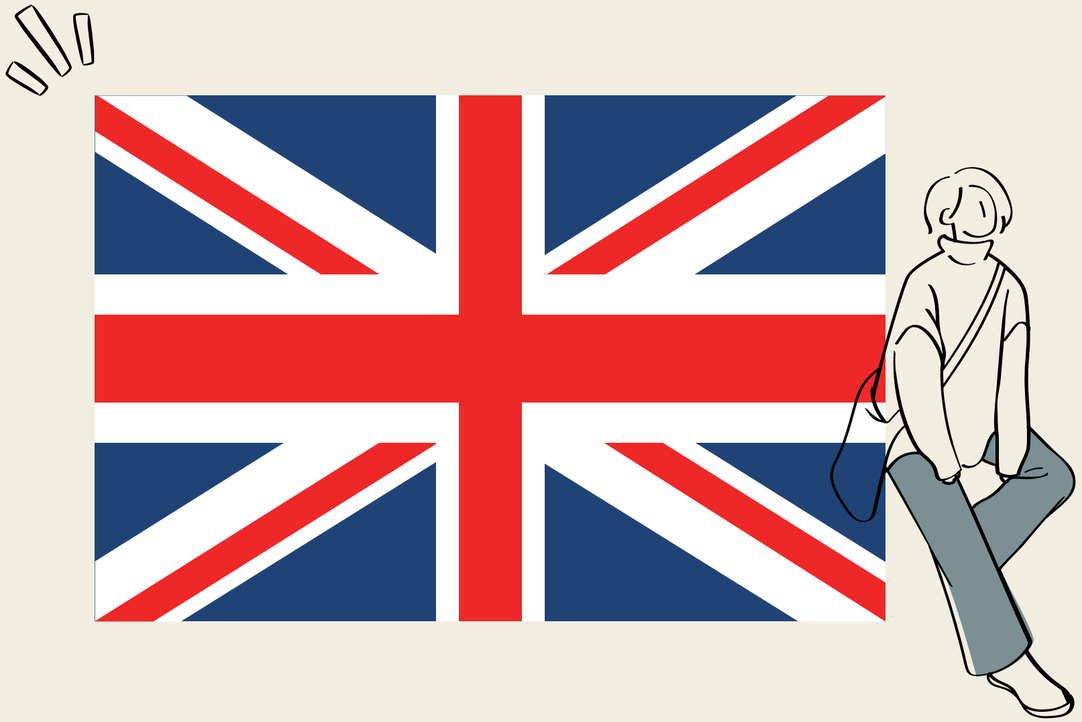
National corpora

Modern linguistic theories

Olga Samieva | LMas-101

The British National Corpus



100 million words of written texts (90%) and transcripts of speech (10%)

– Domain
– Date
– Medium

Free, can be downloaded

Written:
– Books
–Newspapers, magazines and journals
– Electronic (emails, web pages etc)
– Miscellaneous (published and unpublished)

Spoken: face-to-face/phone conversations, speech, meetings

22 million of written and spoken data

Free (via Open ANC), or DVD ($75, from the LDC). Full text access.

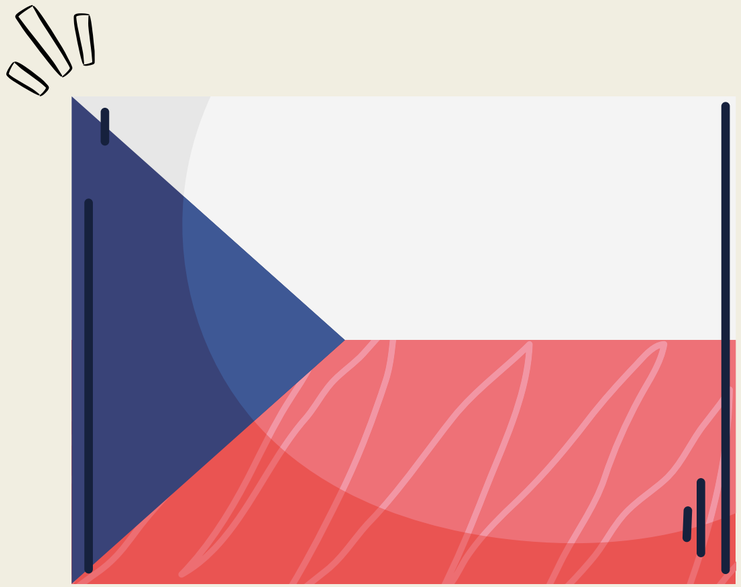
The American National Corpus

The Polish National Corpus

10 million of written and spoken data

Within the framework of PELCRA (Polish and English Language Corpora for Research and Application)

The Czech National Corpus



Synchronous and diachronic

100 million words of the written component of the synchronous section
800,000 words of transcription of authentic spoken language

The Russian Reference Corpus

100 million words of modern Russian

– Modern written texts
– Mid-18th to mid-20th century texts
– Corpus of Spoken Russian

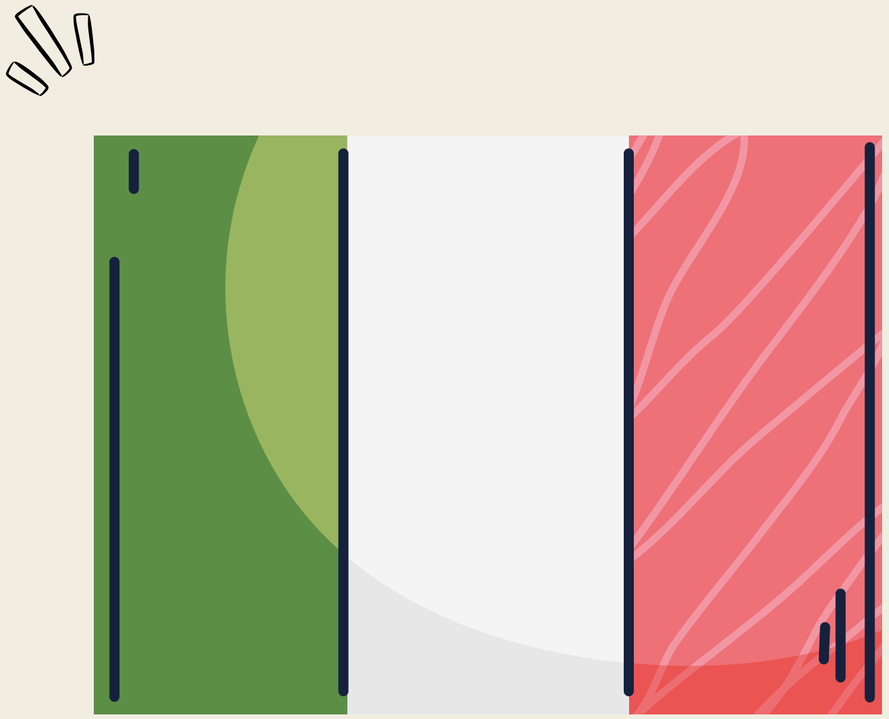
– media (52.7%)
– literature (9.43%)
– scientific texts (13.34%)
– official documents (12.95%)
– informal texts (e.g. electronic forum discussion, 11.58%)

53.7 million words of texts

Free of charge after registration

The Hungarian National Corpus

The CORIS corpus



100 million words of written Italian

– CORIS/CODIS
– DiaCORIS
– BoLC

The Hellenic National Corpus

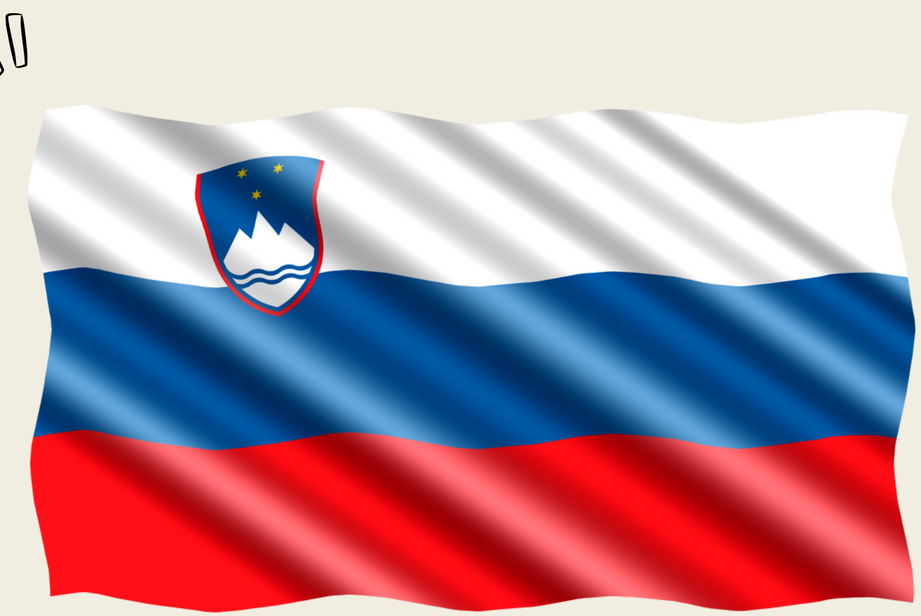
- 32-million-word corpus of written Modern Greek
- Publication media and domains

100-million-word balanced core

–balanced opportunistic subcorpus

The German National Corpus

The Slovak National Corpus



30-million-word corpus

Full and free-of-charge access to the main corpus, subcorpora and other databases after registration

700 million Chinese characters sampled systematically from texts of 1.4 billion characters

– Humanities and social sciences
– Natural sciences
– Miscellaneous
– Newspapers

The MCLC license can be purchased from the National Language Committee of China.

The Modern Chinese Language Corpus

Thank you for your attention!