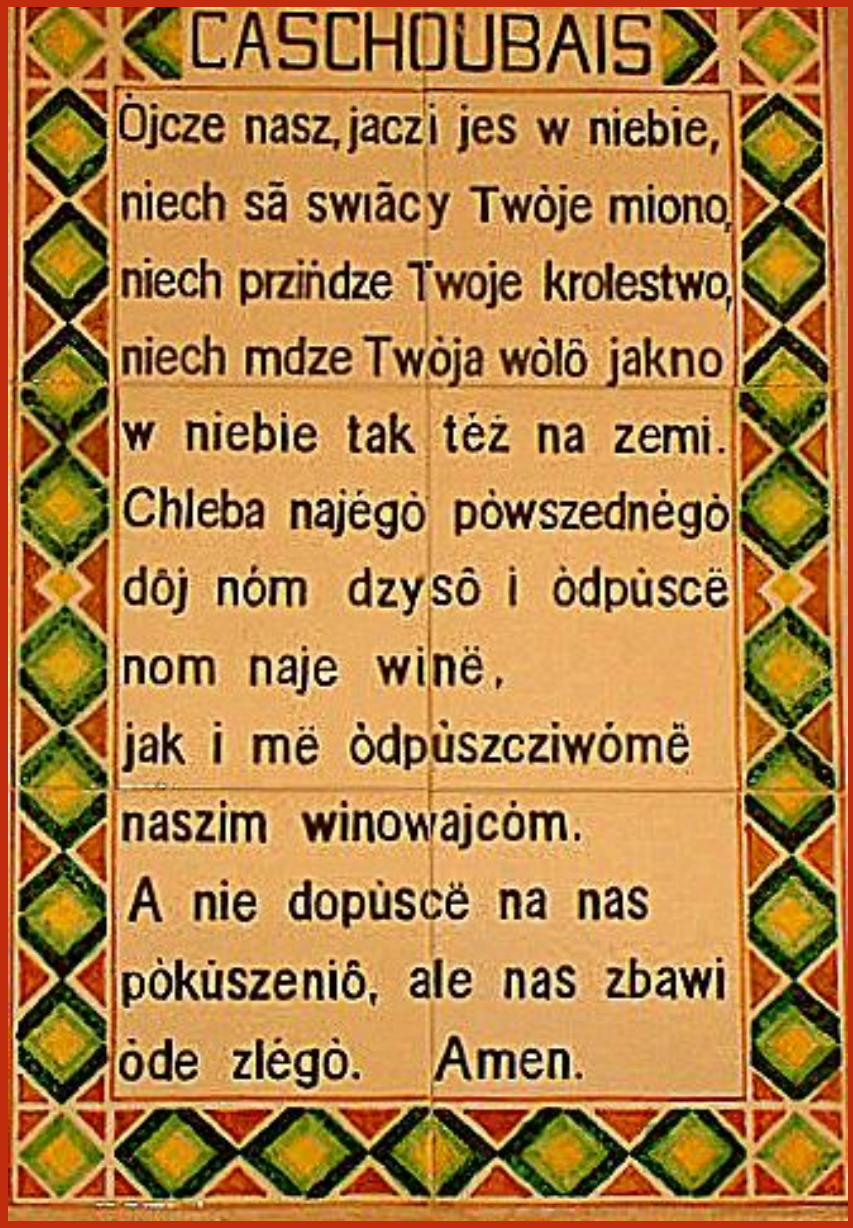
Presented by Darren Nugent & Michael Cronin

CS6361 – Language Engineering & Translation Technology

Kashubian Language Keyboard & Lexical Model

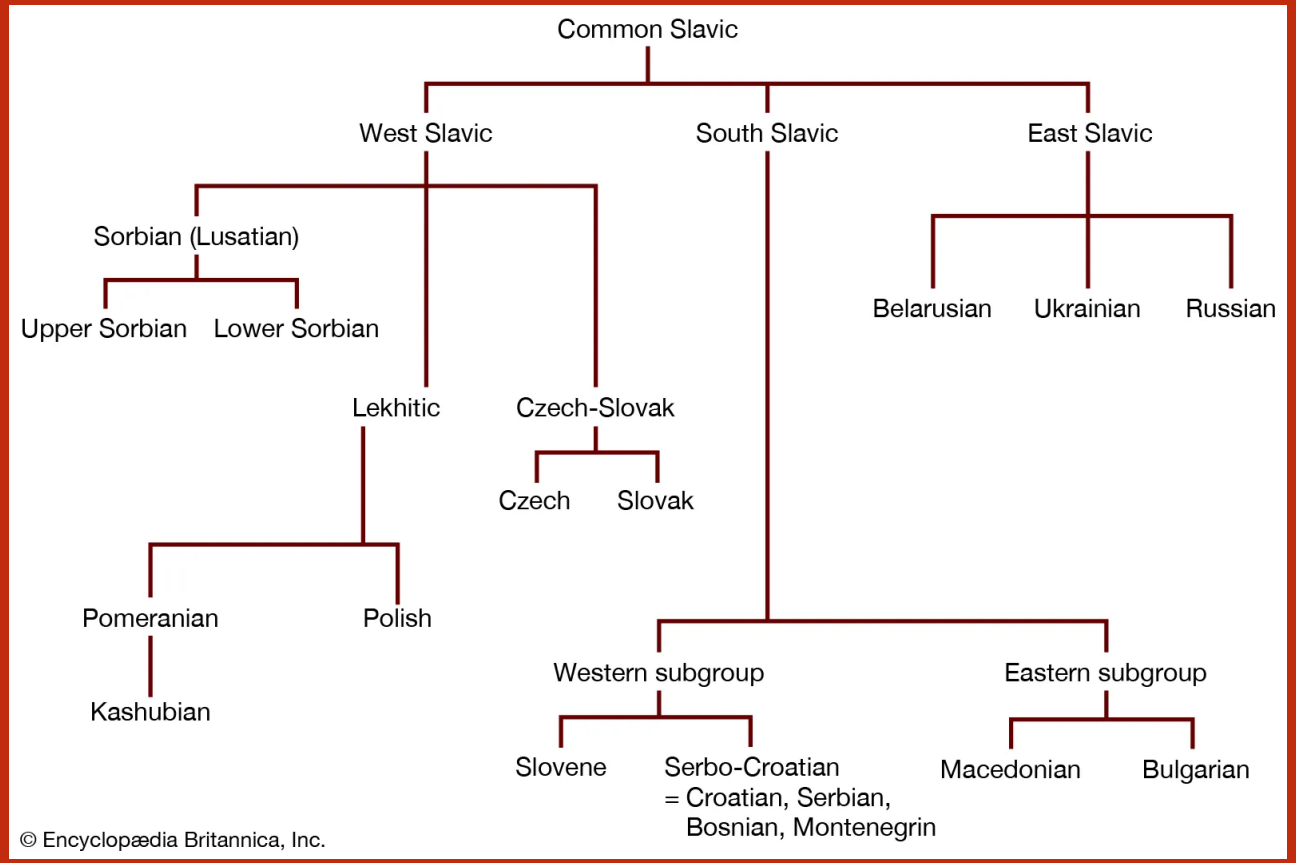


Kashubian: A West Slavic Minority Language

- 100,000-200,000 speakers in Pomerania, Poland
- Recognized regional language

Language characteristics:

- 34-letter Latin-based alphabet
- 11 special diacritics + 6 phonemic digraphs
- Shared orthographic heritage with Polish



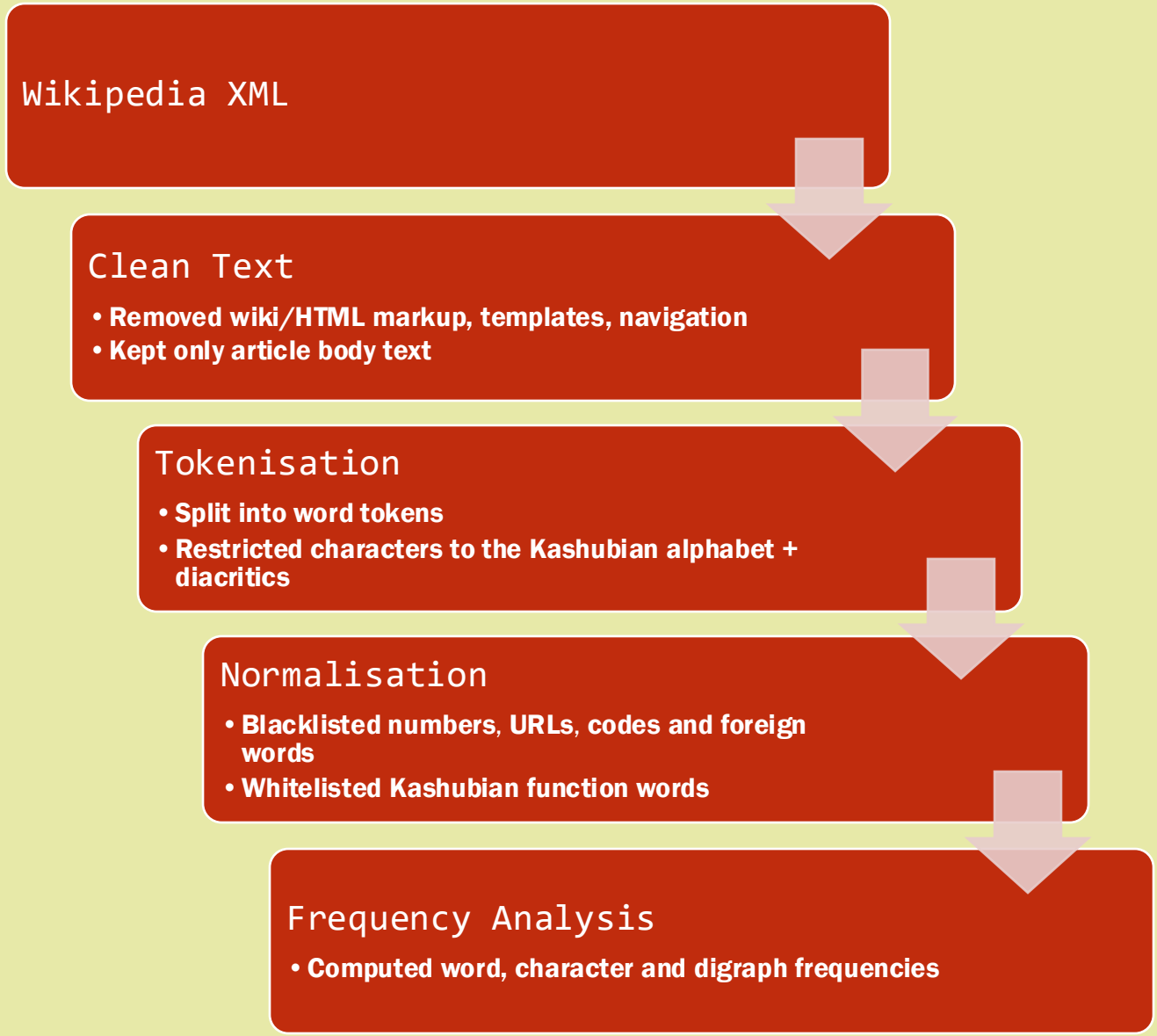
Character Set Identification

- 34-letter Latin-based alphabet
  - 23 standard Latin letters
  - 11 Kashubian-specific diacritics
- 6 phonemic digraphs (single sounds written with two letters)
  - ch, cz, dz, dż, rz, sz
  - Represent distinct consonant phonemes
- Full Unicode (UTF-8) compliance
  - Cross-platform compatibility
  - Keyman Developer integration

Letter	Name (short)
ą	U+0105
ã	U+00E3
é	U+00E9
ë	U+00EB
ł	U+0142
ń	U+0144
ò	U+00F2
ó	U+00F3
ô	U+00F4
ù	U+00F9
ż	U+017C

Corpus Selection & Processing

- **Source:** Kashubian Wikipedia Corpus
- **Corpus Statistics:**
  - 6,933 Articles
  - 84,963 **Unique** Word Forms
  - 498,394 Word Tokens
  - 3.3 Million Characters
- **Kept Stopwords:**
  - Essential for accurate character frequencies
  - High-frequency function words are critical for keyboard design
  - w (3.40%), je (1.63%), na (1.41%)

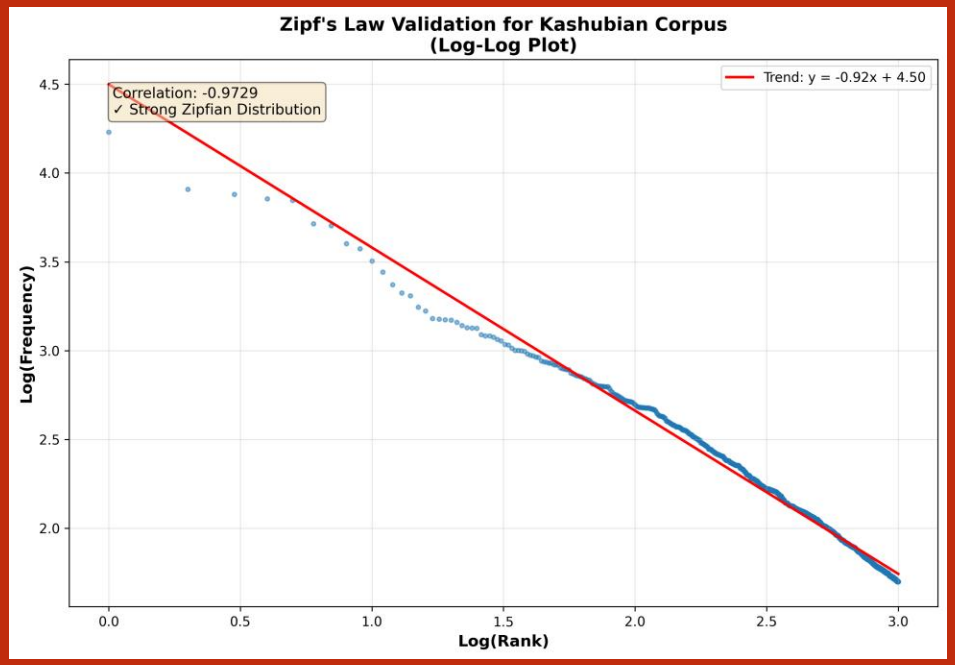


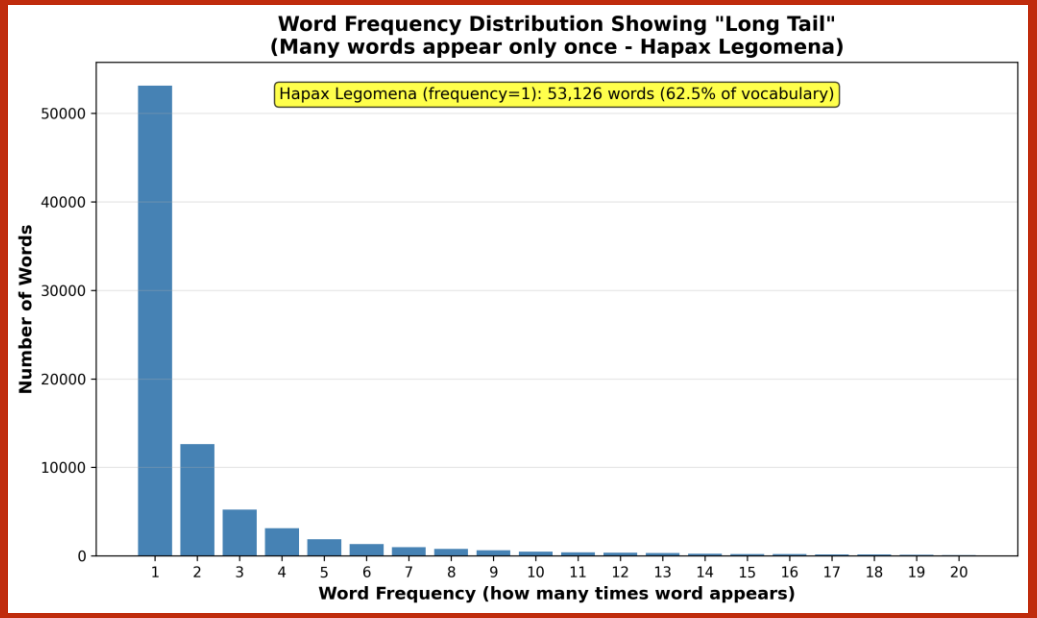
Corpus Quality – Zipf's Law

- **Correlation:** -0.9729
- Expected for strong Zipfian: < -0.85
- Confirms excellent corpus quality
- frequency × rank ≈ constant

VOCABULARY DISTRIBUTION

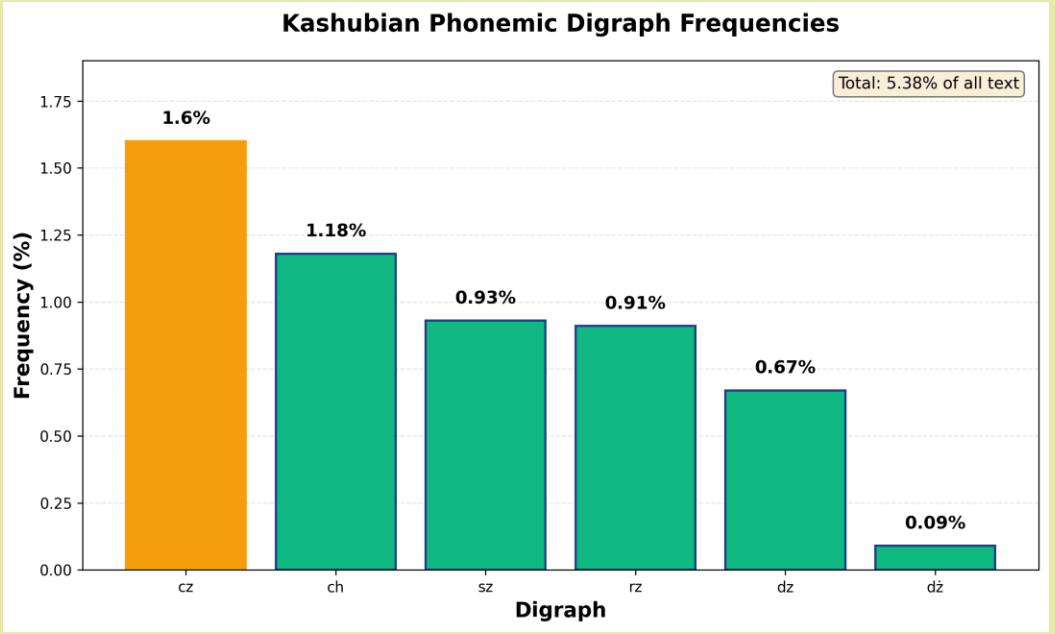
- 62.8% Hapax legomena (appear once)
- 15% Dis legomena (appear twice)
- 22.2% 3 or more occurrences

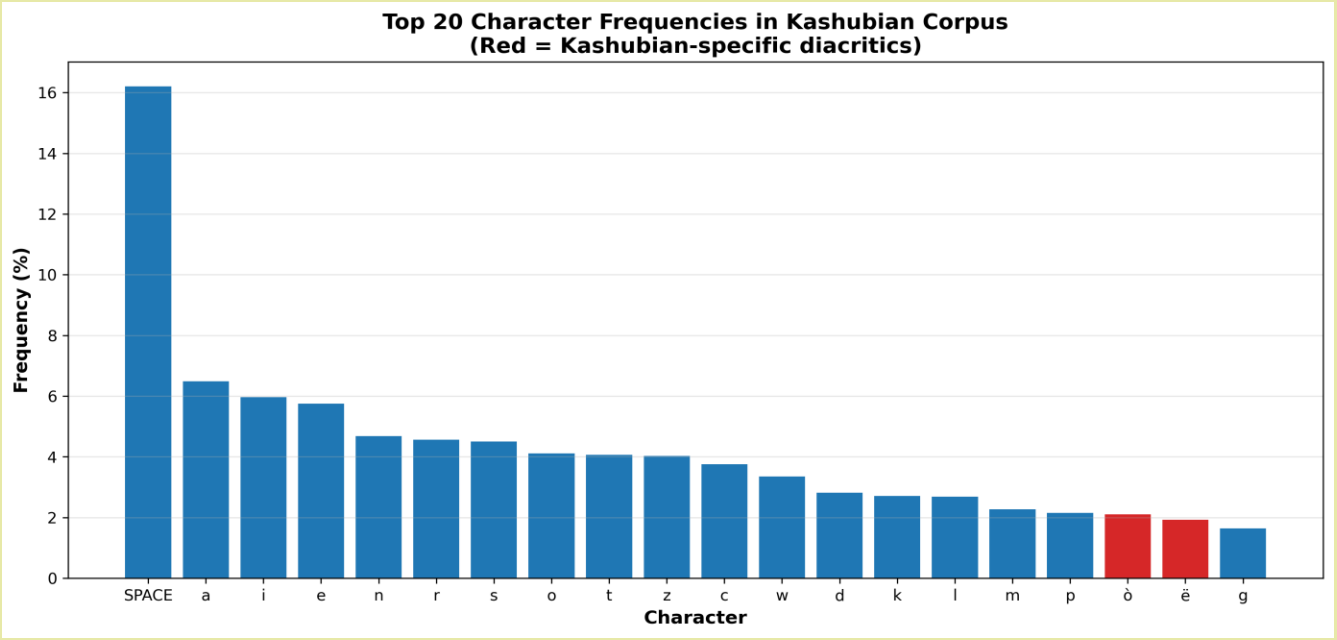




Character Frequency Analysis

- Diacritics:
  - ò ranks #18 (2.11%)
  - ë ranks #19 (1.93%)
  - Both in top 20 most common characters
- Phonemic Digraphs:
  - Cz – 1.6%
  - Ch – 1.18%
  - Sz – 0.93%
  - Rz - 0.91%
  - Dz – 0.67%
  - Dż - 0.09%
  - These 6 phonemic digraphs represent 5.37% of all text





Modified QWERTY

Why QWERTY?

- User familiarity
- Kashubian speakers already know QWERTY
- Lower learning curve than custom layout
- Better cross-platform compatibility
- Proven adoption success for minority languages

Alternatives Considered:

- AZERTY - Not relevant to Poland region
- QWERTZ
- Z/Y swap not helpful for Kashubian
- Dvorak Too unfamiliar, adoption barrier
- Custom - Learning curve prevents adoption

For minority languages, adoption is more beneficial than optimisation!

Desktop Keyboard Design

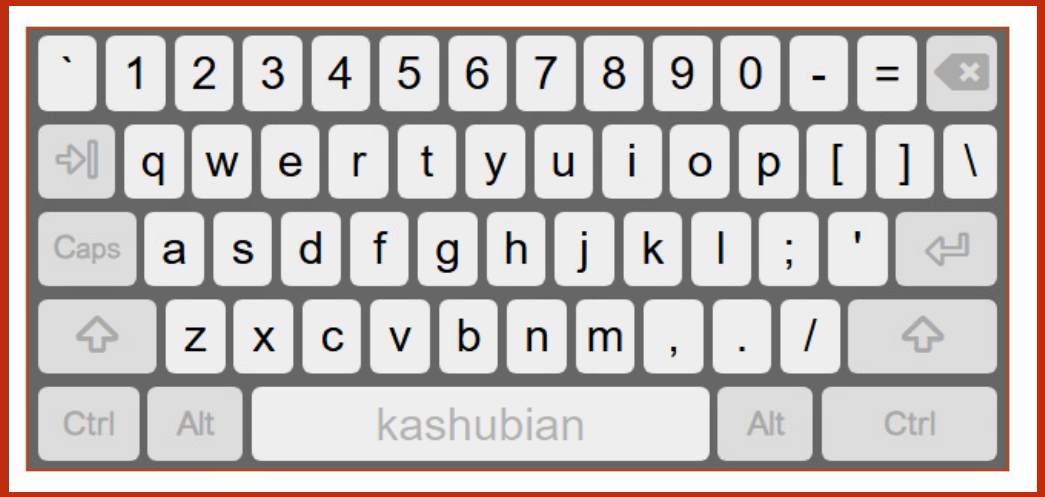
Frequency-Optimised Placement

Character	Frequency	Position	Justification
ò	2.11%	AltGr+O	Prime position
ë	1.93%	AltGr+E	Prime position
ą	0.47%	AltGr+A	Home row
ł	0.89%	AltGr+L	Home row
ń	0.34%	AltGr+N	Accessible

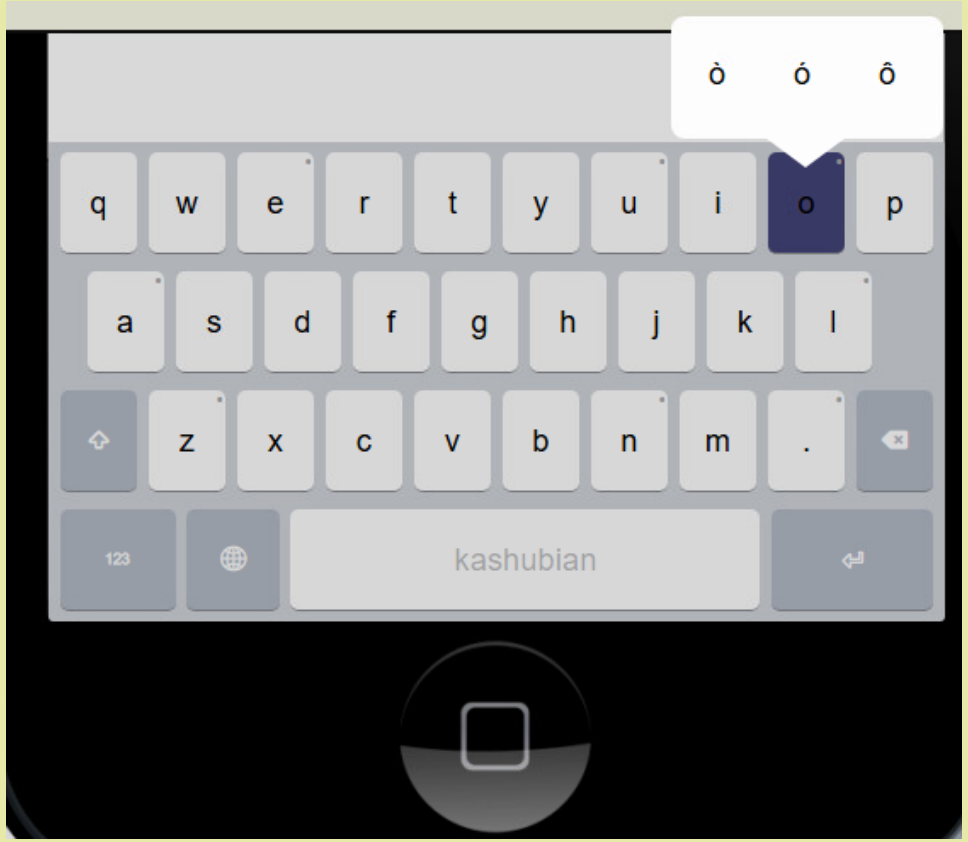
Design Principles:

- Minimise finger travel for frequent characters
- Strong fingers (index, middle) for high-frequency
- Logical grouping (o-diacritics near O key)
- Consistency with Polish where possible

Base Layer: QWERTY



AltGr Layer: Kashubian Diacritics



Long-press	Justification
'o' → ò, ó, ô	ò first: 2.11%
e' → ë, é	ë first: 1.93%
'a' → ą, ã	ą first: 0.47%
'l' → ł	0.89%
'n' → ń	0.34%
'z' → ż	0.44%
'u' → ù	0.60%

Mobile Keyboard Design

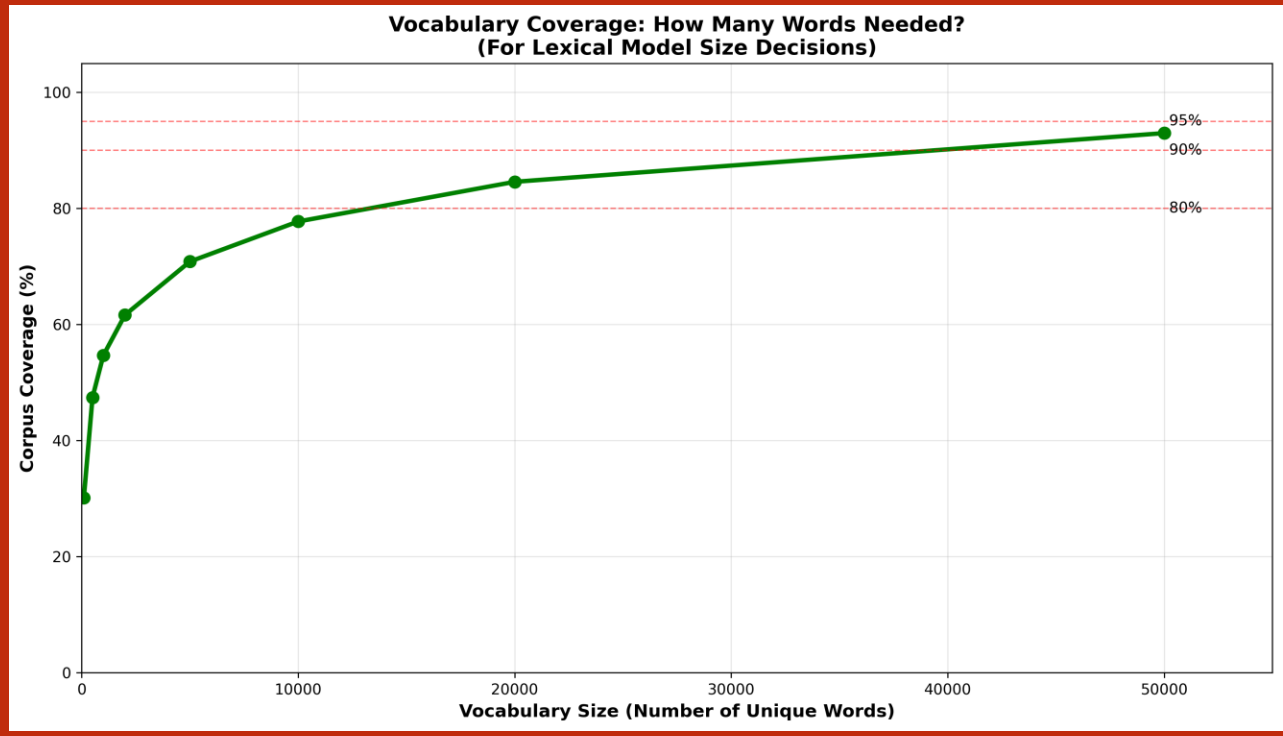
- Users associate diacritics with base
- No extra keyboard layers
- One action to access any diacritic
- Matches smartphone conventions
- Indicators show availability

Lexical Model

Rank	Word	Frequency
1	w	4.16%
2	je	1.99%
3	to	1.85%
4	i	1.76%
5	na	1.72%
6	z	1.24%
7	a	0.98%
8	do	0.92%
9	rok	0.78%
10	są	0.68%

Filtering Decision:
Starting vocabulary: 84,007 unique
Removed hapax legomena: 52,170
Final lexical model: 31,837
Coverage achieved: 89.34%

Rationale:
Two-stage filtering:
(1) Remove single-occurrence words
(2) Systematic blacklist removes English words and markup while protecting Kashubian function words via whitelist



Words	Coverage
1,000	54.7%
10,000	77.8%
31,837	89.34% ← Chose Model
84,007	100.0%

Advanced Modeling Research

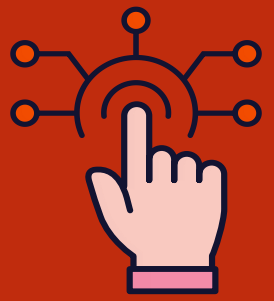
Frequency Model → **N-GRAMS Models** → **Neural Models**

31,837 words 89.34% coverage Ready for Keyman	253,935 bigrams 314,263 trigrams Markov assumption Smoothing: Witten-Bell, Kneser-Ney	Transformers (more data needed)
Limitations: No context No morphology	Capabilities: Context-aware Handles unseen Better Prediction	Requirements: Transfer Learning from Polish Morphological analyzers

1. **Desktop Keyboard**
2. **Mobile Keyboard**
3. **Lexical Model**

All design decisions grounded in extensive corpus analysis

Kashubian severely underrepresented in digital spaces



Brings the digital world to 100k+ Kashubian speakers

Deliverables & Language Preservation

- Wikimedia Foundation. "Kashubian Wikipedia Dump." November 2024. dumps.wikimedia.org/csbwiki/
- Unicode Consortium. "The Unicode Standard, Version 15.0." 2022. unicode.org
- SIL International. "Keyman Developer Documentation." keyman.com/developer/
- Python 3.12 (xml.etree, collections, matplotlib, numpy)
- Jurafsky, D., & Martin, J.H. "Speech and Language Processing" (Chapter 3: N-gram Language Models). Draft of August 24, 2025.
- ISO 639-3 Registration Authority. "Code: csb" iso639-3.sil.org
- Encyclopædia Britannica, Inc. "Slavic languages' family tree." Encyclopædia Britannica. Retrieved November 27, 2025, from https://www.britannica.com/topic/Slavic-languages

References