

Kashubian Keyboard Project - 10-MINUTE PRESENTATION

Complete Script with Timing

Total Time: 10:25

Slides: 14 (includes character set identification & references)

SLIDE 1: Title (20 seconds)

Visual: Lord's Prayer in Kashubian with diacritics highlighted

"Good morning. Today we're presenting our Kashubian keyboard project—a data-driven approach to language digitisation for a low-resource minority language spoken in northern Poland."

SLIDE 2: About Kashubian (35 seconds)

Visual: Map of Pomerania region + sample text with highlighted diacritics

"Kashubian is a West Slavic minority language spoken by approximately 100,000 to 200,000 people in the Pomerania region of northern Poland. It's recognized as a regional language under EU minority language protection. While related to Polish, it's a distinct language with its own orthography—a 34-letter alphabet including 11 special diacritics. As a low-resource language, Kashubian lacks many of the digital tools available for major languages, making systematic keyboard development particularly important."

SLIDE 3: Character Set Identification (40 seconds)

Visual: Complete 34-letter alphabet with Unicode codes for diacritics

"Following systematic methodology, we first identified Kashubian's complete character set. Kashubian uses a 34-letter alphabet: 23 standard Latin letters plus 11 special diacritics shown here with their Unicode code points. We identified this character set through multiple authoritative sources: Kashubian Wikipedia orthography standards, ISO 639-3 language documentation, and Unicode Consortium specifications. Our corpus analysis validated that all 11 diacritics actually appear in real text, confirming our character set is complete and production-ready."

SLIDE 4: Why Kashubian Needs a Keyboard (45 seconds)

Visual: Comparison of Polish keyboard vs Kashubian needs

"Following the systematic methodology from our lectures, we identified five reasons Kashubian needs its own keyboard. Most critically: the most frequent Kashubian diacritic, ó, appears at 2.11% frequency—higher than

the letter 'g' in English—but wasn't easily accessible on any existing keyboard. Polish keyboards have Ł but completely miss ò, ē, and ã. That's like asking you to type English without 'e' or 'a.' This data-driven observation drove our entire design approach."

SLIDE 5: Corpus Selection (40 seconds)

Visual: Wikipedia logo + key statistics

"We used Kashubian Wikipedia as our corpus—6,933 articles, 498,000 words, 85,000 unique vocabulary items. We processed it following the standard NLP pipeline: XML parsing, diacritic-aware tokenization, case normalization. We kept stopwords because they're essential for understanding actual character frequencies. The corpus size is substantial for a minority language and gave us reliable statistics for every design decision."

SLIDE 6: Zipf's Law Validation (50 seconds)

Visual: Log-log plot showing linear relationship

"Before designing anything, we validated corpus quality. Our word frequency distribution follows Zipf's Law with a correlation of -0.9729—that's near-perfect and confirms we're working with high-quality natural language. We found 62.5% hapax legomena—words appearing once—which is exactly the 'long tail' expected in authentic text. This validates that our entire analysis is built on solid data, not noise."

SLIDE 7: Character Frequency Analysis (50 seconds)

Visual: Bar chart of top characters with Kashubian diacritics highlighted

"Here's where data drives design. Character frequency analysis across 3.3 million characters revealed that ò appears at 2.11%—ranking 18th overall, ahead of standard letters. The second Kashubian diacritic, ē, appears at 1.93%, ranking 19th. These two characters alone account for 4% of all text. This data directly determined keyboard placement: ò and ē HAD to be in prime positions. The most common digraph is cz at 1.6%, which informed our ergonomic decisions about sequential key placement."

SLIDE 8: Base Layout Decision (35 seconds)

Visual: QWERTY diagram

"We chose modified QWERTY over custom layouts. Why? User adoption is critical for minority languages. Kashubian speakers already know QWERTY from Polish keyboards. A completely new layout would create barriers that could prevent adoption entirely. We keep QWERTY untouched and add Kashubian diacritics via AltGr layer—familiarity plus functionality."

SLIDE 9: Desktop Keyboard Design (55 seconds)

Visual: Three-layer keyboard diagram with annotations

"Our desktop keyboard has three layers. Base layer: standard QWERTY. Shift layer: capitals. AltGr layer: Kashubian diacritics. Placement is frequency-optimized: AltGr+O gives you ò—the 2.11% character gets the prime position. AltGr+E gives you ë—the 1.93% character in second prime position. Lower-frequency diacritics get less convenient but still logical positions. We applied ergonomic principles: minimize finger travel for frequent characters, use strong fingers for high-frequency operations, maintain consistency with Polish layout where possible."

SLIDE 10: Mobile Keyboard Design (45 seconds)

Visual: Mobile keyboard with long-press indicators

"Mobile uses long-press for diacritics. Long-press 'o' reveals ò, ó, ô—with ò appearing first because of its 2.11% frequency. Long-press 'e' shows ë then é. This is intuitive because users naturally associate diacritics with base letters, and it matches iOS/Android conventions. The mobile keyboard relies heavily on our lexical model for predictions since accessing diacritics requires that extra long-press step."

SLIDE 11: Lexical Model (50 seconds)

Visual: Coverage curve and top words table

"For the lexical model, we analyzed word frequencies. Our 85,000 unique words showed the classic Zipf distribution. We filtered out the 62.5% hapax legomena—single-occurrence words that provide poor prediction value—leaving 31,837 words that give us 89.34% corpus coverage. That's optimal: excellent coverage without bloat. The top 1,000 words alone cover 55% of text. This frequency-based model works well for common words but has limitations—no context awareness, no morphological understanding. Kashubian is morphologically rich like all Slavic languages, but no lemmatization tools exist for it yet."

SLIDE 12: Advanced Modeling Research (50 seconds)

Visual: Model comparison table

"Our frequency model has limitations. Advanced approaches would apply the Markov assumption—predicting words based on previous context—using bigrams or trigrams. For handling unseen n-grams, techniques like Witten-Bell or Kneser-Ney smoothing are standard. Neural models like transformers offer even better performance but are extremely data-hungry. The challenge for low-resource languages is insufficient training data—we have half a million words; GPT needed billions. Potential solutions include transfer learning from multilingual models or leveraging similar languages like Polish. Morphological analyzers would help dramatically but don't exist for Kashubian."

SLIDE 13: Implementation & Impact (50 seconds)

Visual: Deliverables checklist + impact diagram

"We implemented everything in Keyman Developer—desktop keyboard, mobile layout, lexical model. All diacritics are accessible, placement is frequency-optimized, it's ready for deployment.

The broader impact: this keyboard moves Kashubian from partial to good digital support. It enables Wikipedia contributions, social media use, educational content—expanding Kashubian into digital domains which is critical for language vitality. Digital participation is a fundamental right. Our keyboard removes one barrier for a minority language to thrive online.

Thank you—happy to answer questions."

SLIDE 14: References (displayed during Q&A)

Visual: Standard reference list format

DATA SOURCES

- Wikimedia Foundation. Kashubian Wikipedia Dump. dumps.wikimedia.org/csbwiki/
- Unicode Consortium. Unicode Standard 15.0. [unicode.org](https://www.unicode.org)

TOOLS & DEVELOPMENT

- SIL International. Keyman Developer Documentation. keyman.com
- Python 3.12 (xml.etree, collections, matplotlib, numpy)

ACADEMIC FRAMEWORK

- CS6361 Lecture Materials (Weeks 4, 5, 7, 8)
- Zipf, G.K. (1949). Human Behavior and the Principle of Least Effort
- Witten-Bell & Kneser-Ney smoothing techniques

LANGUAGE RESOURCES

- Kashubian orthography standards (post-2000 standardization)
- ISO 639-3 code documentation (csb)

This slide appears during Q&A period - not part of the 10-minute timing.

TIMING BREAKDOWN

Slide	Topic	Time
1	Title	0:20
2	About Kashubian	0:35
3	Character Set	0:40

Slide	Topic	Time
4	Why Needs Keyboard	0:45
5	Corpus Selection	0:40
6	Zipf's Law	0:50
7	Character Frequency	0:50
8	Base Layout	0:35
9	Desktop Design	0:55
10	Mobile Design	0:45
11	Lexical Model	0:50
12	Advanced Research	0:50
13	Implementation & Impact	0:50
14	References	(Q&A)
TOTAL		10:25

Buffer: 35 seconds to hit "under 11 minutes"

TIMING CHECKPOINTS

Use these during practice to track your pace:

Checkpoint	Target Time
After Slide 2 (About Kashubian)	0:55
After Slide 4 (Why Keyboard)	2:20
After Slide 7 (Character Freq)	5:25
After Slide 10 (Mobile)	7:45
After Slide 13 (Impact)	10:25

If running long: Cut examples from Slides 11-12

If running short: Expand design rationale in Slides 9-10

KEY CONTENT RETAINED

- ✓ Language introduction (NEW)
- ✓ Character set identification with sources (NEW - REQUIRED)
- ✓ Systematic keyboard development methodology
- ✓ Zipf's Law validation (Week 4)
- ✓ Hapax legomena discussion (Week 4)
- ✓ Data-driven design decisions

- Frequency-optimized placement
 - Desktop and mobile designs
 - Lexical model coverage analysis
 - Markov assumption & smoothing techniques (Week 4)
 - Morphological challenges
 - Impact on language preservation
 - References (REQUIRED)
-

DELIVERY TIPS

Speak clearly but efficiently - You have dense content in 10 minutes

Use visuals - Let charts/diagrams carry technical details

Practice transitions - "Moving to character frequency..." saves time

Key numbers to emphasize:

- **34 letters, 11 diacritics** (Slide 3)
- **2.11%** (ò frequency)
- **-0.9729** (Zipf correlation)
- **62.5%** (hapax legomena)
- **89.34%** (coverage)

Key phrases to use:

- "Data-driven design"
 - "Frequency-optimized"
 - "Systematic methodology"
 - "Authoritative sources" (for character set)
 - "Fundamental right to digital participation"
-

Q&A PREPARATION (Common Questions)

Q: Why not use Polish keyboard? "Polish has Ł but completely lacks ò, è, and ã—the three most frequent Kashubian-specific characters at 2.11%, 1.93%, and 0.60% respectively. Using Polish would force users to struggle with characters they need in every sentence."

Q: How did you identify the character set? "We used three authoritative sources: Kashubian Wikipedia orthography standards which reflect the post-2000 standardization, ISO 639-3 language documentation, and

Unicode Consortium specifications. Our corpus analysis then validated that all 11 diacritics actually appear in authentic text."

Q: Did you test with users? "Not yet—user testing is our next phase. However, our design is grounded in 500,000 words of native speaker text from Wikipedia, so it reflects authentic usage patterns."

Q: What about advanced models? "Bigram or trigram models with Witten-Bell or Kneser-Ney smoothing are the natural next step. Longer-term, transfer learning from Polish could bring sophisticated predictions, since both are West Slavic languages with similar morphology."

Q: How does this help language preservation? "Digital presence equals language vitality. Languages used in fewer domains are more vulnerable to extinction. Our keyboard expands Kashubian into digital space—social media, education, work—especially for young people who live online."

Q: Can this work for other languages? "Absolutely. Our methodology—systematic character set identification, corpus analysis, frequency-based design—is generalizable. Any language with a reasonable corpus and Unicode support can follow this approach."

PROJECT REQUIREMENTS COVERAGE

Requirement	Our Coverage	Slide(s)
1. Language selection + rationale	<input checked="" type="checkbox"/> Excellent	2, 4
2. Corpus identification	<input checked="" type="checkbox"/> Excellent	5
3. Character set + how identified	<input checked="" type="checkbox"/> Excellent	3
4. Base layout decision	<input checked="" type="checkbox"/> Good	8
5. Frequency analysis + approach	<input checked="" type="checkbox"/> Excellent	6, 7
6. Desktop keyboard design	<input checked="" type="checkbox"/> Excellent	9
6. Mobile keyboard design	<input checked="" type="checkbox"/> Excellent	10
7. Lexical model building	<input checked="" type="checkbox"/> Excellent	11
8. Advanced model research	<input checked="" type="checkbox"/> Good	12
9. Referenced throughout	<input checked="" type="checkbox"/> Yes	All
10. Final reference slide	<input checked="" type="checkbox"/> Yes	14

All requirements met! ✓

MODULE INTEGRATION CHECKLIST

Week 4 Concepts:

- Zipf's Law validation (Slide 6)
- Hapax legomena discussion (Slide 6, 11)

- N-gram analysis (Slide 7)
- Markov assumption (Slide 12)
- Smoothing techniques (Slide 12)

Week 5 Concepts:

- Text preprocessing pipeline (Slide 5)
- Tokenization with diacritics (Slide 5)
- Case normalization (Slide 5)
- Stopword decisions (Slide 5)
- Corpus quality assessment (Slide 6)

Week 7 Concepts:

- Language digitisation importance (Slide 2, 13)
- ISO codes (Slide 3)
- Unicode encoding (Slide 3)
- Character set identification (Slide 3)
- Morphological complexity (Slide 11, 12)

Week 8 Concepts:

- Why languages need keyboards (Slide 4)
- Systematic keyboard development (All slides)
- Ergonomic principles (Slide 9, 10)
- Digital participation rights (Slide 13)

Complete module integration! ✓

SUCCESS METRICS

You'll know you succeeded if:

- ✓ Presentation finishes 10:00-11:00
- ✓ All 13 content slides covered
- ✓ Character set identification explained
- ✓ Key numbers stated clearly
- ✓ Methodology explained coherently
- ✓ Questions answered confidently

- ✓ You maintained composure
 - ✓ References displayed during Q&A
-

FINAL CHECKLIST - DAY OF PRESENTATION

Physical Materials:

- Laptop fully charged
- USB backup of slides
- Phone backup of slides
- Printed script
- Water bottle
- Watch or timer visible

Digital Files:

- Slide deck tested
- All fonts render correctly
- Unicode characters display (ò, ä, à, etc.)
- All images/charts work
- References slide ready

Mental Preparation:

- Know key numbers (34, 11, 2.11%, -0.9729, 89.34%)
 - Practice character set explanation
 - Have Q&A answers ready
 - Positive mindset
 - Deep breaths
-

PRACTICE RECOMMENDATIONS:

- 1. Time yourself** on each slide with stopwatch
- 2. Mark checkpoints** at 2:20, 5:25, 7:45
- 3. Practice character set slide** - it's new content
- 4. Rehearse Q&A** - especially character set identification
- 5. Record yourself** - watch for pace and clarity
- 6. Get feedback** from someone unfamiliar with project

You're ready to deliver an excellent, requirements-complete presentation!

Good luck! 