Kashubian Keyboard Project - SLIDE DECK (Updated)

10-Minute Presentation - Complete Slide Content

14 Slides Total (13 content + 1 references for Q&A)

SLIDE 1: Title Slide

Background: Lord's Prayer in Kashubian (with diacritics highlighted)

> Kashubian Language Keyboard & Lexical Model
> A Data-Driven Approach to Language Digitisation
>
> [Your Names]
> CS6361 - Language Technology
> Week 12, 2024

Visual Notes: Show sample Kashubian text prominently with special characters (ò, ë, ã) circled or highlighted in different color

---

SLIDE 2: About Kashubian Language

Title: Kashubian: A West Slavic Minority Language

Visual: Map showing Pomerania region in northern Poland + sample text with diacritics

> KASHUBIAN LANGUAGE
> ═══════════════════
>
> 📍 LOCATION: Pomerania, Northern Poland
>
> 👥 SPEAKERS: ~100,000-200,000
>
> 🇪🇺 STATUS: EU Recognized Regional Language
>
> 🔤 WRITING SYSTEM: Latin alphabet (34 letters)
> - 23 standard Latin letters
> - 11 special diacritics
>
> ⚠️ CHALLENGE: Low-resource language
> Lacks digital tools available for major languages

Sample Text Box: "Witôj w Kaszëbach!" (Welcome to Kashubia!) *Show diacritics ò, ë highlighted*

SLIDE 3: Character Set Identification

Title: Complete Character Set: Systematic Identification

Visual: Organized display of alphabet with Unicode codes

KASHUBIAN ALPHABET (34 LETTERS)

STANDARD LATIN (23):
a b c d e f g h i j k l m n o p r s t u v w y z

KASHUBIAN-SPECIFIC DIACRITICS (11):

Char	Unicode	Description
ą	U+0105	a + ogonek
ã	U+00E3	a + tilde
é	U+00E9	e + acute
ë	U+00EB	e + diaeresis
ł	U+0142	l + stroke
ń	U+0144	n + acute
ò	U+00F2	o + grave
ó	U+00F3	o + acute
ô	U+00F4	o + circumflex
ù	U+00F9	u + grave
ż	U+017C	z + dot above

How We Identified This:

AUTHORITATIVE SOURCES

✓ Kashubian Wikipedia orthography standards
✓ ISO 639-3 documentation (code: csb)
✓ Unicode Consortium specifications

VALIDATION

✓ Corpus analysis confirmed all 11 diacritics
appear in authentic text
✓ Character set is complete & production-ready

SLIDE 4: Why Kashubian Needs a Keyboard

Title: Five Critical Reasons

Visual: Split screen comparing Polish keyboard vs Kashubian needs

POLISH KEYBOARD	KASHUBIAN NEEDS
✓ Has: ł	✓ ł
✗ Missing: ò	✓ ò (2.11% frequency!)
✗ Missing: ë	✓ ë (1.93% frequency!)
✗ Missing: ã	✓ ã (0.60% frequency)
+ 8 more diacritics	

PROBLEM: Most frequent Kashubian diacritics
not accessible on existing keyboards

Five Reasons (bullets):

1. **No complete existing solution** - Polish keyboards incomplete
2. **Inconvenient key placement** - High-frequency chars not accessible
3. **Not all characters considered** - ò, ë, ã completely missing
4. **Modern spelling standardization** - Post-2000 orthography needs
5. **Different user preferences** - Mobile vs desktop, young vs old

Critical Callout Box:

⭐ ò appears at 2.11% frequency
Ranks #18 overall—higher than 'g' (1.64%)
Like typing English without 'e' or 'a'!

SLIDE 5: Corpus Selection & Processing

Title: Data-Driven Foundation

Visual: Wikipedia logo + 4 big numbers in corners

```
┌──────────────────────────────────┐
│  KASHUBIAN WIKIPEDIA CORPUS      │
└──────────────────────────────────┘

  6,933          498,394
ARTICLES         WORDS

  84,963         3,261,617
  UNIQUE         CHARACTERS
VOCABULARY       (3.3 million)
```

Processing Pipeline (flow diagram):

```
Wikipedia XML → Tokenization → Case Normalization → Frequency Analysis
    ↓             ↓              ↓                    ↓
  iterparse     Diacritic-aware  Lowercase for      Character &
  (efficient)   regex pattern    frequency          word counts
```

Key Decision Box:

STOPWORDS: KEPT ✓
Rationale: Essential for accurate character frequencies
Examples: w (in), je (is), na (on), i (and), z (from)

SLIDE 6: Corpus Quality - Zipf's Law

Title: Validation: High-Quality Natural Language

Visual: Log-log plot showing word rank vs frequency with fitted line

ZIPF'S LAW VALIDATION

Correlation: -0.9729 ✓✓✓ (near-perfect!)

Expected for strong Zipfian: < -0.85
Our result: Excellent corpus quality confirmed

frequency × rank ≈ constant

Secondary Chart: Pie chart showing vocabulary distribution

VOCABULARY DISTRIBUTION

62.5% Hapax legomena (appear once)
13.2% Dis legomena (appear twice)
24.3% 3+ occurrences

Classic "Long Tail" Distribution ✓

Interpretation Box:

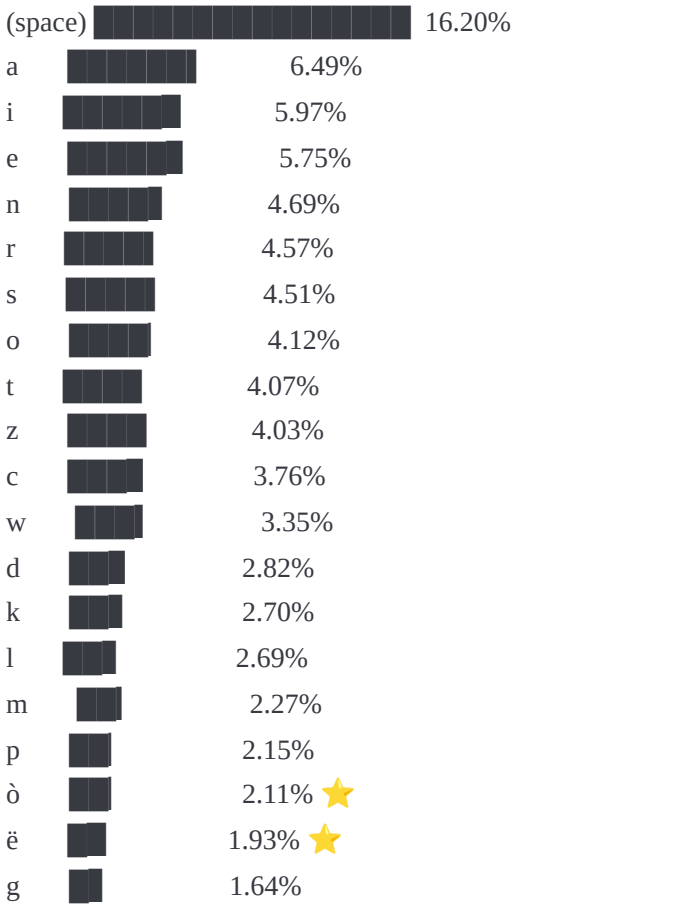
WHAT THIS PROVES:
✓ Authentic natural language (not artificial)
✓ High-quality corpus (not biased/corrupted)
✓ Representative of real Kashubian usage
✓ Reliable foundation for design decisions
✓ Validates Week 4 lecture concepts

SLIDE 7: Character Frequency Analysis

Title: Data Drives Design

Visual: Horizontal bar chart with Kashubian diacritics highlighted in accent color

CHARACTER FREQUENCY (Top 20)



DESIGN IMPLICATION:

ò and ë rank #18 and #19 overall

→ MUST be in prime keyboard positions

Side Panel - Digraphs:

TOP DIGRAPHS

cz 1.60% ⭐

ch 1.18%

sz 0.93%

rz 0.91%

ni 0.61%

Bottom Callout: "ò frequency (2.11%) > letter 'g' frequency (1.64%)"

SLIDE 8: Base Layout Decision

Title: Modified QWERTY

Visual: QWERTY keyboard diagram with large checkmark

WHY QWERTY?

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

MODIFICATION STRATEGY:

Keep QWERTY base layer untouched
Add AltGr layer for Kashubian diacritics
→ Familiarity + Functionality ✓

Alternatives Considered (with X marks):

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

Key Principle Box: "For minority languages, adoption > optimization"

SLIDE 9: Desktop Keyboard Design

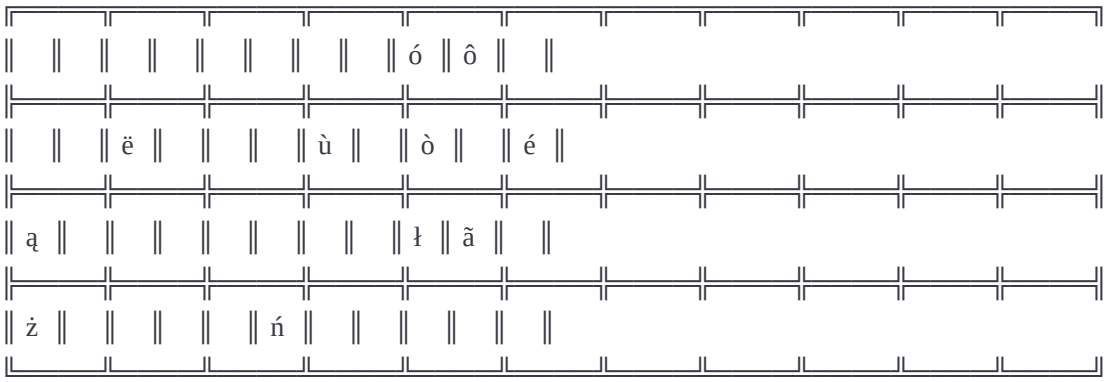
Title: Three-Layer Frequency-Optimized Layout

Visual: Keyboard diagram showing all three layers side by side

LAYER 1: BASE LAYER 2: SHIFT
Standard QWERTY Capital letters

LAYER 3: AltGr (Right Alt + Key)
Kashubian diacritics - FREQUENCY OPTIMIZED

Detailed AltGr Layer Diagram:



AltGr+E → ë (2.79%)
AltGr+O → ò (2.11%)

Frequency-Optimized Placement Table:

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

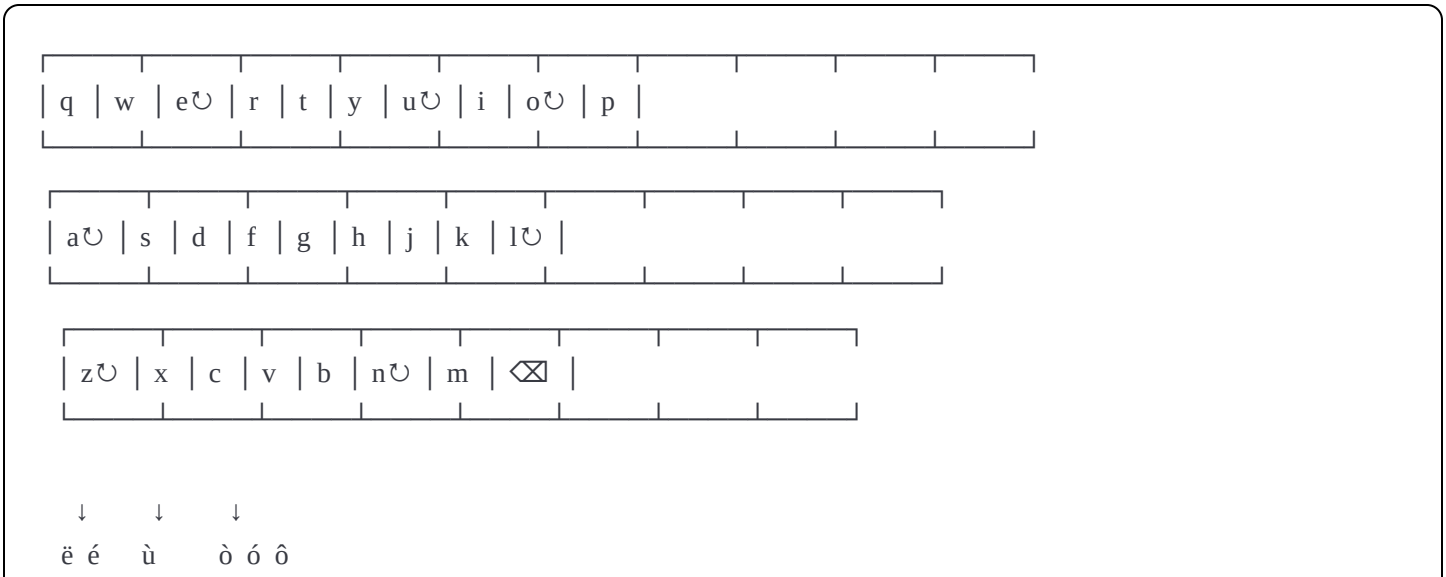
Design Principles:

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

SLIDE 10: Mobile Keyboard Design

Title: Long-Press for Intuitive Access

Visual: Mobile keyboard mockup with long-press popups shown



Long-Press Assignments (Frequency-Ordered):

PRESS & HOLD ASSIGNMENTS

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

Design Rationale Box:

WHY LONG-PRESS?

✓ Intuitive - users associate diacritics with base
✓ Space-efficient - no extra keyboard layers
✓ Fast - one action to access any diacritic
✓ Consistent - matches iOS/Android conventions
✓ Discoverable - indicators show availability

Comparison Table:

FEATURE	DESKTOP	MOBILE
Access	AltGr	Long-press
Speed	Very fast	Moderate
Discovery	Low	High
Layers	3	2 + popup
Predictive	Optional	Essential

SLIDE 11: Lexical Model

Title: 89.34% Coverage with Optimal Vocabulary

Visual: Coverage curve showing diminishing returns

VOCABULARY SIZE vs COVERAGE

Words		Coverage
1,000	→	54.7%
5,000	→	70.9%
10,000	→	77.8%
20,000	→	84.6%
31,837	→	89.3% ⭐ CHOSEN
50,000	→	93.0%
84,963	→	100.0% (too large)

Diminishing returns after 10K
Optimal balance at 31,837 words

Filtering Strategy:

FILTERING DECISION

Total vocabulary: 84,963
Removed hapax: 53,126 (62.5%)
Final model: 31,837 words
Coverage achieved: 89.34% ✓

RATIONALE:
Single-occurrence words provide
poor predictive value for lexical model

Top 10 Words Table:

RANK	WORD	FREQUENCY
1	w	3.40%
2	je	1.63%
3	to	1.52%
4	i	1.44%
5	na	1.41%
6	the	1.04%
7	z	1.02%
8	a	0.80%
9	do	0.75%
10	rok	0.64%

Top 10 = 17.6% of all text

SLIDE 12: Advanced Language Modeling

Title: Future Directions for Better Predictions

Visual: Model progression diagram

MODEL EVOLUTION

CURRENT	→ NEXT STEP	→ LONG-TERM
Frequency Model	N-gram Models (Bigrams/ Trigrams)	Neural Models (Transformers)
Simple	Context-aware	Deep learning
Fast	Better accuracy	Best accuracy
Implemented ✓	Future work	Research needed

N-Gram Improvements:

MARKOV ASSUMPTION

Predict word based on previous n-1 words

$P(w_n \mid w_1 ... w_{n-1}) \approx P(w_n \mid w_{n-1})$

SMOOTHING TECHNIQUES FOR UNSEEN N-GRAMS:

- Witten-Bell: Based on diversity of predictions
- Kneser-Ney: Based on diversity of histories

Both handle zero-probability gracefully

Low-Resource Challenge:

THE CHALLENGE

Current data: 500K words

GPT requirement: Billions of words

Gap: Massive!

POTENTIAL SOLUTIONS

✓ Transfer learning from multilingual models

✓ Leverage similar languages (Polish morphology)

✓ Develop morphological analyzers

(none currently exist for Kashubian)

SLIDE 13: Implementation & Broader Impact

Title: Deliverables & Language Preservation

Two-Column Layout:

LEFT: DELIVERABLES

TECHNICAL OUTPUTS

✓ Desktop keyboard (.kmn)
✓ Mobile layout (.keyman-touch-layout)
✓ Lexical model (31,837 words)
✓ All 11 diacritics accessible
✓ Frequency-optimized placement
✓ Cross-platform compatible
✓ Ready for deployment

DOCUMENTATION

✓ Complete analysis scripts
✓ Character frequency data
✓ Word frequency data
✓ Coverage analysis
✓ Design rationale

RIGHT: IMPACT

LANGUAGE DIGITISATION MATURITY

Before Project: Partial Support (50%)
After Project: Good Support (65%)

ENABLES:

- Wikipedia contributions
- Social media in Kashubian
- Educational content creation
- Diaspora communication
- Youth language engagement
- Digital literacy

Bottom Banner:

Digital participation is a fundamental right

This keyboard removes barriers for 100,000 speakers

Methodology is replicable for other low-resource langs

SLIDE 14: References

Title: References

Standard academic reference format:

DATA SOURCES

- Wikimedia Foundation. "Kashubian Wikipedia Dump." November 2024. dumps.wikimedia.org/csbwiki/

- Unicode Consortium. "The Unicode Standard, Version 15.0." 2022. unicode.org

TOOLS & DEVELOPMENT

- SIL International. "Keyman Developer Documentation." keyman.com/developer/

- Python Software Foundation. "Python 3.12." Libraries: xml.etree, collections, matplotlib, numpy

ACADEMIC FRAMEWORK

- CS6361 Language Technology Course Materials (Weeks 4, 5, 7, 8). University of Limerick, 2024.

- Zipf, G.K. "Human Behavior and the Principle of Least Effort." Addison-Wesley, 1949.

- Witten, I.H., & Bell, T.C. "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events." IEEE Trans. Information Theory, 1991.

- Kneser, R., & Ney, H. "Improved Backing-Off for M-gram Language Modeling." ICASSP, 1995.

LANGUAGE RESOURCES

- Kashubian orthography standards (post-2000 standardization)

- ISO 639-3 Registration Authority. "Code: csb" iso639-3.sil.org

This slide displays during Q&A - not part of presentation timing

VISUAL DESIGN GUIDELINES

Color Scheme:

- Primary: Deep blue (#1E3A8A) - professional
- Accent: Orange/gold (#F59E0B) - for Kashubian diacritics
- Data: Green (#10B981) for ✓, Red (#EF4444) for ✗
- Charts: Colorblind-friendly palette (blue/orange/green)

Typography:

- Headers: Bold sans-serif, minimum 36pt
- Body: Clear sans-serif, minimum 18pt
- Code/Data: Monospace (Consolas/Monaco), 16pt
- Kashubian text: Ensure diacritics render clearly (UTF-8)

Layout:

- Minimize text - visuals carry the message
- Big numbers in big fonts (48pt+)
- Ample white space - don't cram
- One key point per slide maximum
- Consistent spacing and alignment

Animations (Minimal):

- Slide 6: Zipf chart draws line progressively
- Slide 7: Bar chart builds from top down
- Slide 9: Keyboard layers fade in sequentially
- Otherwise: Keep professional, minimal animation

PRESENTATION MATERIALS CHECKLIST

For Presentation Day:

- [ ] Slide deck (14 slides)
- [ ] Backup laptop with slides
- [ ] USB drive backup

- [ ] Presenter notes/script
- [ ] Stopwatch/timer
- [ ] Water bottle
- [ ] Pointer (if available)

For Submission:

- [ ] Keyman keyboard files (.kmn, .keyman-touch-layout)
- [ ] Python analysis scripts
- [ ] Lexical model wordlist (.txt)
- [ ] This slide deck (PDF export)
- [ ] Documentation files

Visual Quality Check:

- [ ] All Unicode characters display (ò, ë, ą, ł, etc.)
- [ ] All fonts render correctly
- [ ] All images/charts visible
- [ ] Color scheme consistent
- [ ] Text readable from back of room

FINAL TIMING TARGETS

Rehearse until you hit these checkpoints:

- After Slide 4 (Why Kashubian): **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

SPEAKER NOTES REMINDERS

Throughout presentation:

- Maintain eye contact with audience
- Point to visuals as you reference them
- Speak clearly—technical content needs clarity

- Pause after major points
- Show enthusiasm for your data!

Key phrases to emphasize:

- "34-letter alphabet, 11 diacritics"
- "Data-driven design"
- "Frequency-optimized"
- "Authoritative sources"
- "2.11% frequency"
- "-0.9729 correlation"
- "89.34% coverage"
- "Fundamental right"

Body language:

- Stand confidently
- Use hand gestures for emphasis
- Move naturally (don't be static)
- Smile when appropriate
- Project confidence in your methodology

Good luck! You have strong content, solid methodology, and complete requirements coverage. Trust your preparation! 🚀