# The Use of Alphanumeric Symbols in Slovene Tweets

DAFNE MARKO

4th Conference on CMC and Social Media Corpora
for the Humanities
27–28 September 2016
Faculty of Arts, Ljubljana

# Outline

- Goal
- Theoretical background
- Dataset and methodology
- Results
- Qualitative Analysis
- Conclusion

# Goal

- Identify the most frequently used words with alphanumeric symbols in Slovene tweets
- Comparison among other CMC genres + the Kres corpus (standard Slovene) → we expect to find no words with alphanumeric symbols in the Kres corpus, proving they are a CMC-specific feature
- Comparison according to user gender, user type, text standardess
- Analysis of the most frequently used numerals

# Theoretical background

- Different expressions for the phenomenon described:
  - (alphanumeric) rebus writing (Halmetoja, 2013; Danet and Herring, 2007)
  - complex abbreviation (Filipan-Žignić et al., 2012)
  - textism (Grace et al., 2012; Bushnell et al., 2011)
  - rebus-like potential of words (Crystal, 2001)
  - letter/number homophone (Bieswanger, 2006; Kirsten Torrado, 2014; Frehner, 2008; Thurlow, 2003; Alkawas, 2011, etc.)

- Two functions:
  - Word-shortening strategy
  - Way of creative writing → "the way of writing is as important as the content" (Kirsten Torrado, 2014)

# Theoretical background

- Major characteristic of letter/number homophones:
  - the **pronunciation** of numerals is identical with letters or parts of words, enabling them to replace a letter or letter sequences

  - Focus mostly on the pronunciation, but not on the **graphical appearance** of numerals

    **b4** for “before” vs. **g33k** for “geek”

# Theoretical background

- Words with alphanumeric symbols identified in Slovene text messages and e-mails (Mihelizza, 2008; Dobrovoljc, 2008; Logar, 2006): ju3 = “jutri”, pr8 = “prosim”, 5er = “Peter”, 1x = “enkrat” mi2 = “midva”
- No research on words with numerals used graphically

# Dataset and Methodology

- For our research, two corpora were used:
  - **the JANES v0.4 corpus →** a large corpus of Slovene tweets, forum posts, blog entries, comments on news articles and on Wikipedia pages and users (over 175 million words)
  - **the Kres corpus →** a collection of standard written Slovene with a balanced genre structure (nearly 100 million words)

- Focus on the biggest subcorpus → Twitter posts written in Slovene (altogether 90.180.337 words from 7.503.199 different Twitter posts)

# Dataset and Methodology

- data extraction with the concordancer SketchEngine
- employing **CQL expressions →** numeral(s) + letter(s); letter(s) + numeral(s) + letter(s); letter(s) + numeral(s)
- **frequency lists** for each position of numerals

- irrelevant results were manually selected and excluded from the list → proper names/part of a proper names, chemical symbols, units of measurement (e.g., *A4*, *24ur*, *CO2*, *C4*, *TEŠ6*, *m2*, etc.)

# Results

- No results for numerals at the beginning of the word → problem with tokenization!

- **Numeral at the end of the word**
  - after excluding irrelevant results, 27 different tokens with 15 different lemmas were found
  - relative frequency = 33.1 per million tokens
  - 6 English words: *hi5/Hi5*; *tr00/Tr00/TR00*; *gr8/Gr8*; *str8*; *h8/H8*; *sk8*
  - 4 Slovene pronouns: *mi2/Mi2/MI2*; *vi2/Vi2*; *mi3/Mi3*; *me2/Me2*

# Results

| Token | Abs. freq. |
|---|---|
| Ju3 | 1173 |
| Mi2 | 593 |
| mi2 | 371 |
| ju3 | 337 |
| s5* | 292 |
| MI2 | 119 |
| vi2 | 110 |
| hi5 | 97 |
| tr00 | 77 |
| zju3 | 50 |
| Hi5 | 47 |
| na1 | 36 |
| gr8 | 36 |

| Token | Abs. freq. |
|---|---|
| Tr00 | 31 |
| Mi3 | 31 |
| me2 | 27 |
| str8 | 26 |
| Vi2 | 20 |
| Gr8 | 17 |
| Me2 | 11 |
| h8 | 11 |
| u3 | 10 |
| sk8 | 7 |
| Zju3 | 5 |
| mi3 | 5 |
| H8 | 3 |

*S5 excluded from the list – used exclusively in the proper name *Galaxy S5*

# Results

- **Numeral in the middle of the word**
  - after excluding irrelevant results, 117 different tokens with 50 different lemmas were found
  - relative frequency = 9.97 per million tokens
  - the list of different words with numerals appearing in the middle of the word is significantly longer, whereas the relative frequency in much lower
  - majority of English words → preposition “to” substituted by number 2 (e.g., *B2B*, *p2p*, *coffee2go*, *up2date*, etc.)

# Results

| Token | Abs. freq. |
|---|---|
| B2B/b2b | 205/41 |
| w00t/W00t | 66/39 |
| d00h/d0h/D0h/ d000h | 51/48/26/4 |
| pr0n/Pr0n | 49/6 |
| g33k/ g33ki/g33kov/ g33ka/G33k | 35/9/6/5/4 |
| na1x | 30 |
| n00b/n00be | 24/4 |
| B2C | 21 |

| Token | Abs. freq. |
|---|---|
| s3ksi/S3ksi | 19/4 |
| p2p/P2P | 19/18 |
| B4B | 19 |
| p0rn | 18 |
| Za1x | 13 |
| mi3je | 12 |
| še1x | 11 |
| ju3šnji/ju3snji/Ju3šnji/ju3šnjega/ju3snjem | 11/4/4/3/3 |

# Results

- **The use of alphanumeric symbols according to user type**
  - strong tendency of **private users** to incorporate such writing into their tweets
  - private users: **70%**; corporate users: 30%

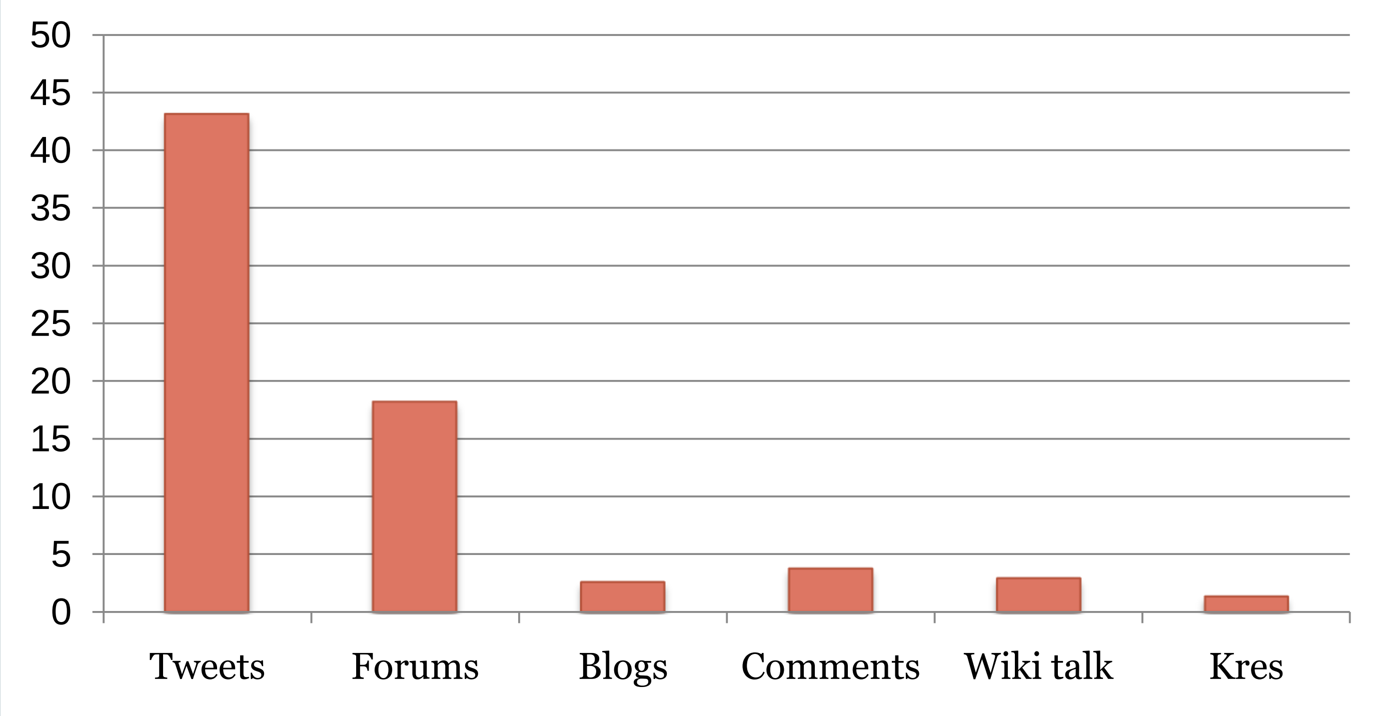
- **The use of alphanumeric symbols according to user gender**
  - words with alphanumeric symbols is far **more frequent among male users**
  - male users: **80%**; female users: 20%

# Results

- **The use of alphanumeric symbols according to level of text standardness**
  - comparison of all 9 possibilities of text standardness – from L1T1 to L3T3
  - Words with alphabetic and numeric symbols most frequently used in tweets annotated as very non-standard (**L3T3**) or linguistically very non-standard and technically slightly non-standard (**L3T2**).

# Results

- Comparison of CMC genres (tweets, forum posts, blog entries, comments on news articles, Wiki talk) and the Kres corpus



# Results

- The Kres corpus
  - A total of 12 different examples – 10 of them with numeral in the middle of the word (e.g., *cig4ni*, *za1x*, *pr0n*), one with numeral ending a word (*ju3*), and one with numeral starting a word (*4ever*)
  - all of these examples were found in the texts obtained from the web pages and from the computer gaming magazine *Joker*

# Qualitative analysis

- In the JANES corpus, 8 numerals were identified: 0, 1, 2, 3, 4, 5, 7, and 8; most frequent ones: 2, 3, 8, and 0

| Numeral | Interpretation | Example |
|---|---|---|
| 1 | “ena”<br>“i” | *na1* = “na ena”<br>*BRA71L* = “Brazil” |
| 2 | “dva”<br>“dve”<br>“to” | *mi2* = “midva”<br>*me2* = “medve”<br>*up2date* = “up to date” |
| 3 | “tri”<br><br>“e” | *ju3 = “jutri”*<br>*s3njam = “strinjam”*<br>*g33k = “geek”* |
| 4 | “for”<br>“a” | *t4t = “training for trainers”*<br>*G4ME = “game”* |

# Qualitative analysis

| Numeral | Interpretation | Example |
|---|---|---|
| 5 | “pet”<br>“five” | *s5 = “spet”*<br>*hi5 = “high five”* |
| 7 | “z” | *BRA71L = “Brazil”* |
| 8 | “eat”<br>“aight”<br>“ate” | *gr8 = “great”*<br>*str8 = “straight”*<br>*h8 = “hate”*<br>*l8r = “later”* |
| 0 | “o” | *n00b = “noob”*<br>*p0rn = “porn”*<br>*w00p = “woop”* |

# Qualitative analysis

## Phonetic vs. graphic function of numerals

- **Phonetic**: the pronunciation of numerals is identical with a letter or sequence of letters, e.g. *s5* = “spet”
- **Graphic**: the graphic appearance of numerals is similar to the substituted letter or string of letters, e.g. G4ME = “GAME”
- Most of the numerals at the end of the words are used phonetically (ju3, mi2, gr8); the only exception:

  tr00 → troo → tru: → “true”
- Most of the numerals in the middle of the word are used graphically (s3ksi, d00h, w00t); exceptions: ju3šnji, mi3je

# Conclusion

- more than 60 Slovene and English words with alphanumeric symbols in Slovene tweets
- characteristic for CMC, especially microtexts (Twitter and forum posts)
- The same numeral can be used phonetically or graphically

# References

- Alkawas, S. (2011). *Textisms: The Pragmatic Evolution among Students in Lebanon and its Effect on English Essay Writing*. Master Thesis, Lebanese American University.
- Baron, S. (2008). *Always On: Language in an Online and Mobile World*. Oxford University Press, Oxford.
- Bieswanger, M. (2006). 2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space- and time-saving strategies in English and German text messages. In Hallett, T., Floyd, S., Oshima, S. and Shield, A. (Eds.), *Texas Linguistics Forum Vol. 50*, Austin.
- Bushnell, C., Kemp, N. and Heritage Martin, F. (2011). Text-messaging practices and links to general spelling skills: A study of Australian children. In *Australian Journal of Educational & Developmental Psychology*. Vol 11, pp. 27–38.
- Crystal, D. (2001). *Language and the Internet*. Cambridge, Cambridge University Press.  
  Danet, B. and Herring, S. (Eds.). (2007). *The*
- *Multilingual Internet*. *Language, Culture, and Communication Online*. Oxford University Press, Oxford.
- Denby, L. (2010). *The Language of Twitter: Linguistic Innovation and Character Limitation in Short Messaging*. Undergraduate dissertation, University of Leeds.
- Dobrovoljc, H. (2008). Jezik v e-poš  tnih sporoč  ilih in vpraš  anja sodobne normativistike. In Koš  uta, M. (Ed.), *Slovenš  č  ina med kulturami*, Slavistič  no druš  tvo Slovenije, Celovec, pp. 295–314.
- Elizondo, J. (2011). *Not 2 Cryptic 2 DCode: Paralinguistic Restitution, Deletion, and Non-standard Orthography in Text Messages*. Ph.D. thesis, Swarthmore College.

# References

- Filipan-Ž ignić, B., Velič ki, D. and Sobo, K. (2012). SMS communication – Croatian SMS language features as compared with those in German and English Speaking Countries. In *Revija za elementarno izobraž evanje*, š t. 1. Pedagoš ka fakulteta, Maribor.
- Fiš er, D., Erjavec, T. and Ljubeš ic´, N. (2016). Janes v0.4: korpus slovenskih spletnih uporabniš kih vsebin. *Slovenš č ina 2.0* (to appear).
- Frehner, C. (2008). *Email, SMS, MMS: The Linguistic Creativity of Asynchronous Discourse in the New Media Age*. Peter Lang.
- Gouws, S., Metzler, D., Cai, C. and Hovy, C. (2011). Contextual bearing on linguistic variation in social media. In *Proceedings of the workshop on language in social media* (*LSM 2011*), pp. 20–29. http://aclweb.org/anthology/W/W11/W11-0704.pdf.
- Grace, A., Kemp, N., Martin, F. H. and Parrila, R. (2012). Undergraduates' use of text messaging language: Effects of country and collection method. In *Writing Systems Research*. Taylor & Francis Online.
- Halmetoja, T. (2013). *Gender-Reated Variation in CMC Language: A Study of Three Linguistic Features on Twitter*. BA thesis, Gö teborgs Universitet.
- Kadir, Z. A., Maros, M. and Hamid, B. A. (2012). Linguistic Features in the Online Discussion Forums. In *International Journal of Social Science and Humanity*, Vol. 2, No. 3, May 2012, pp. 276–281.
- Kirsten Torrado, U. (2014). Development of SMS language from 2000 to 2010. In Cougnon, L. and Fairon, C. (Eds.), *SMS communication: A linguistic approach*. Benjamins Current Topics.
- Kul, M. (2007). Phonology in text messages. In *Poznań Studies in Contemporary Linguistics 43(2)*, pp. 43–57.

# References

- Ljubeš ić, N., Fiš er, D., Erjavec, T., Č ibej, J., Marko, D., Pollak, S. and Š krjanec, I. (2015). Predicting the level of text standardness in user-generated content. In *Proceedings*, pp. 371–378, Hissar: [s.n.]. http://lml.bas.bg/ranlp2015/docs/RANLP_main.pdf.
- Logar Berginc, N., Grč ar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š . and Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za druž bene vede.
- Logar, N. and Smith, J. (trans.). (2006). Stilno zaznamovane nove tvorjenke: tipologija = Stylistically marked new derivates: a typology. In Vidovicˇ-Muha, A. (Ed.), *Slovensko jezikoslovje danes*, Slavistič no druš tvo Slovenije, Ljubljana, pp. 87–101.
- Michelizza, M. (2008). Jezik SMS-jev in SMS-komunikacija. In *Jezikoslovni zapiski: zbornik Inš tituta za slovenski jezik Frana Ramovš a*, Inš titut za slovenski jezik Frana Ramovš a ZRC SAZU, Ljubljana, pp. 151–166.
- Moseley, N. (2013). *Using word and phrase abbreviation patterns to extract age from Twitter microtexts*. Thesis, Rochester Institute of Technology.
- Sherblom-Woodward, B. (2002). *Hackers, Gamers and Lamers: The Use of l33t in the Computer Sub-Culture*. http://www.swarthmore.edu/SocSci/Linguistics/papers /2003/sherblom - woodward.pdf.
- Thurlow, C. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. In
- *Discourse Analysis Online*, Sheffield.

Thank you!