The role of phonotactics in gender assignment in Polish

Marcin Kilarski and Paula Orzechowska
kilarski@amu.edu.pl, paulao@ifa.amu.edu.pl

Contents

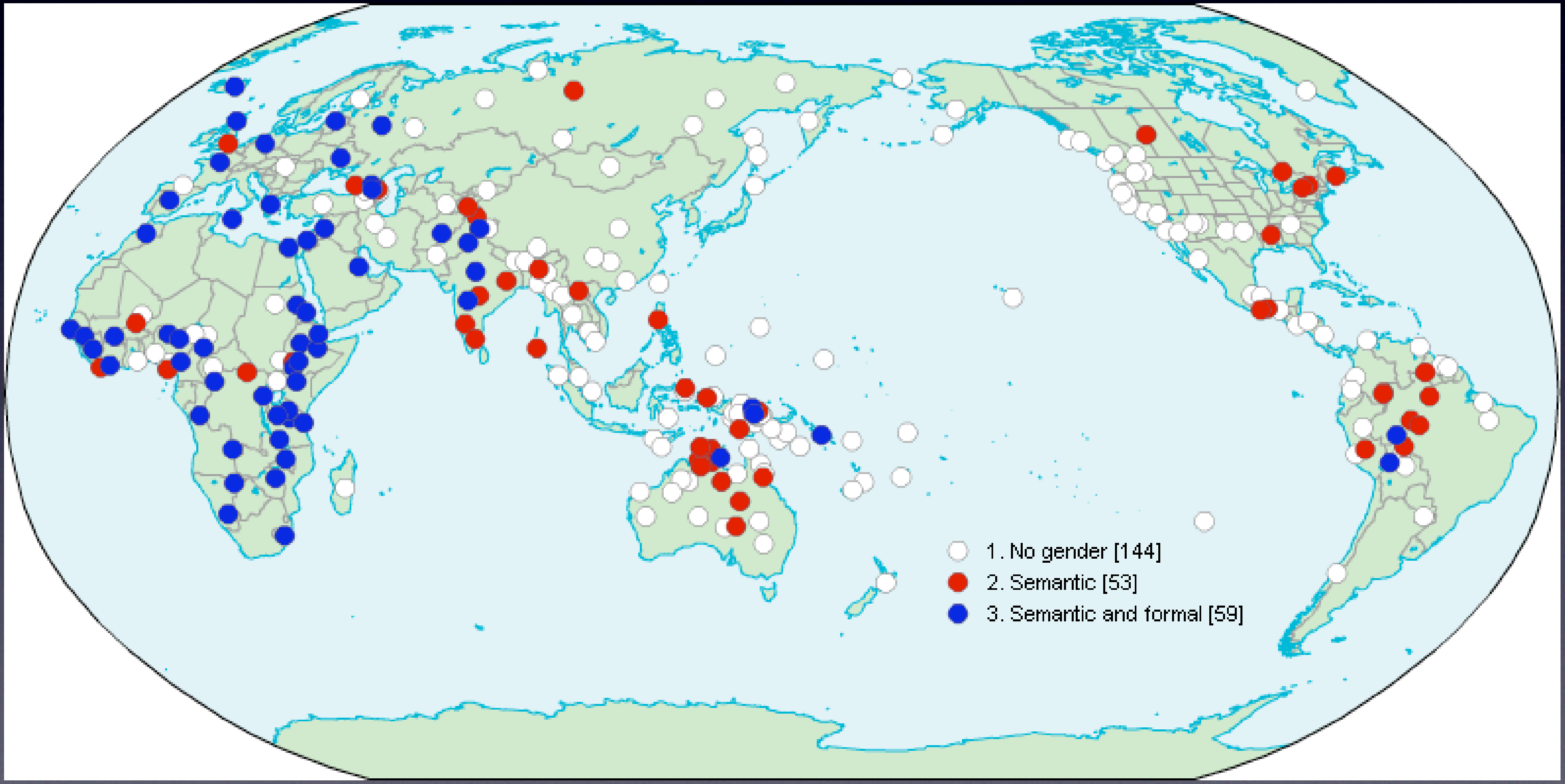
- overview of gender assignment criteria
- Polish: gender assignment; phonotactics
- description of corpus
- role of selected phonotactic criteria
  - size and type of final sequence, tokens
  - correlations with morphological criteria and etymology
- implications for morphology and typology

Gender assignment

- assignment criteria: semantic and formal
  - formal: phonological and morphological (inflectional, derivational)
- phonological criteria: gender established on the basis of a single form (Corbett 1991)
- interplay of criteria (overlap or conflict)
- role of phonological shape in gender and declension

Gender assignment

Formal criteria



Gender assignment

Phonological criteria

- final or initial sequences of phones
    - German: consonant cluster principle; final clusters /-(C)+ f, ç, x + t/ as f. (Köpcke & Zubin)
    - French: backward processing: /ɛzɔ̃ sjɔ̃ zjɔ̃ ʒjɔ̃ tjɔ̃/ as f., others in /ɔ̃/ as m. (Tucker et al.)
    - Godié (Kru): type of final vowel (front : central : back)
- suprasegmentals
    - Qafar (Cushitic): m. (all in -C plus those with non-final accent) : f. (final accented vowel)

Semantic and formal criteria in Polish

- semantic: masc. (male sex-differentiables + residue), fem. (female sex-differentiables + residue) and neut. (residue)
- phonological (word-final phonemes):
  - /k g x/ (*chemik* ‘chemist’ m.); /ɛ/ (*pole* ‘field’ n.)
- morphological (inflection or derivation):
  - m. nouns inflected like f. and ending in -*a* (*poeta* ‘poet’)
  - n. nouns in -*o* (*babsko* ‘woman, augm., pej.’)

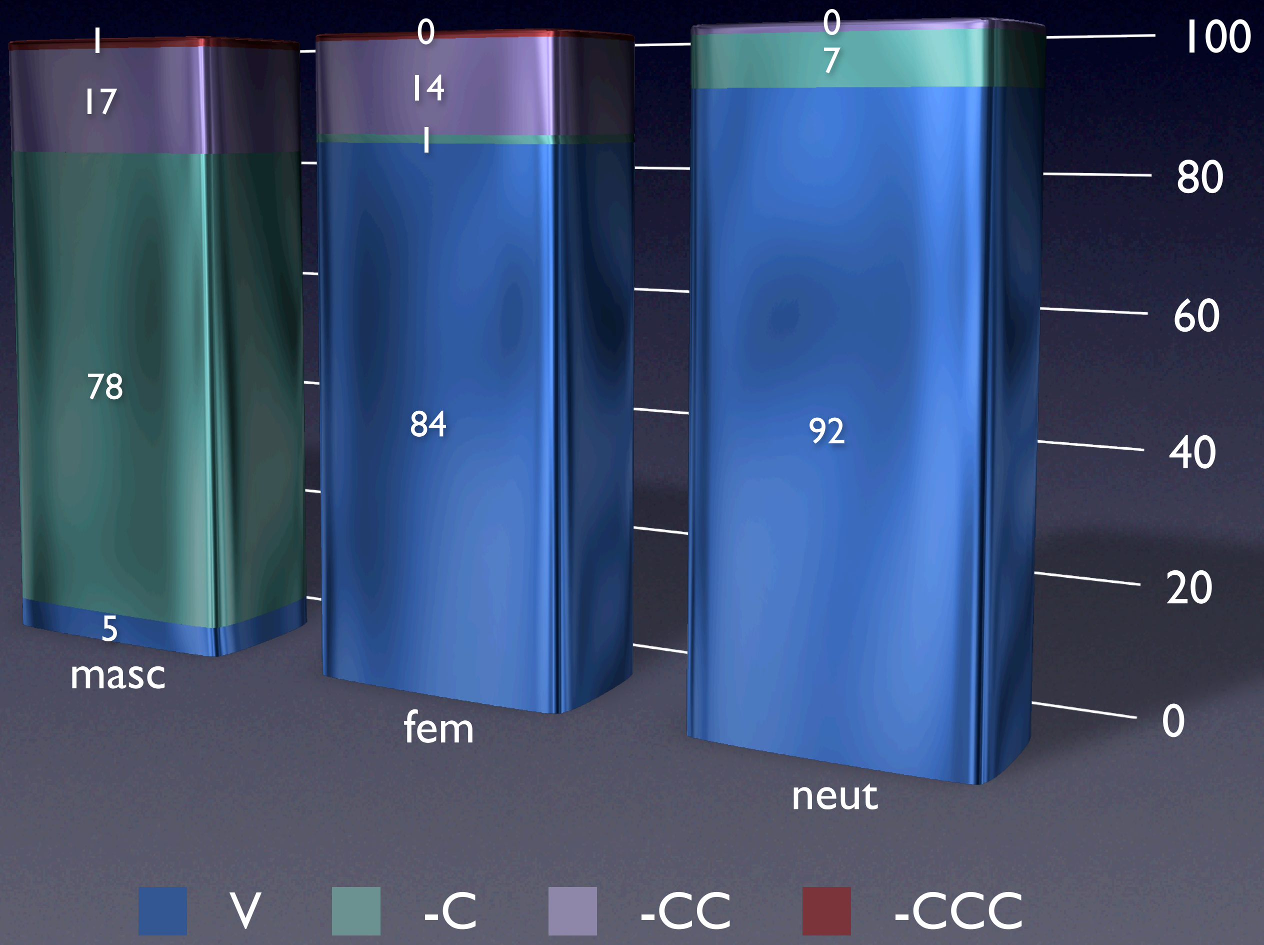
Polish phonotactics

- consonantal language, 5.5:1 C-V ratio, c. 80% C in the inventory
- 5-C sequences (morphonotactic) vs. 3-C (lexical) word-finally
  - *skąpstw+o* ‘stinginess’ --> /skompstf/ (gen.pl.) vs. *mistrz* ‘master’ --> /mistʃ/
- <ą ę> before obstruents
  - *sęp* --> /sɛmp/ ‘vulture’, *łabędź* --> /wabɛɲʥ/ ‘swan’, *kęs* --> /kɛũ̯s/ ‘bite’

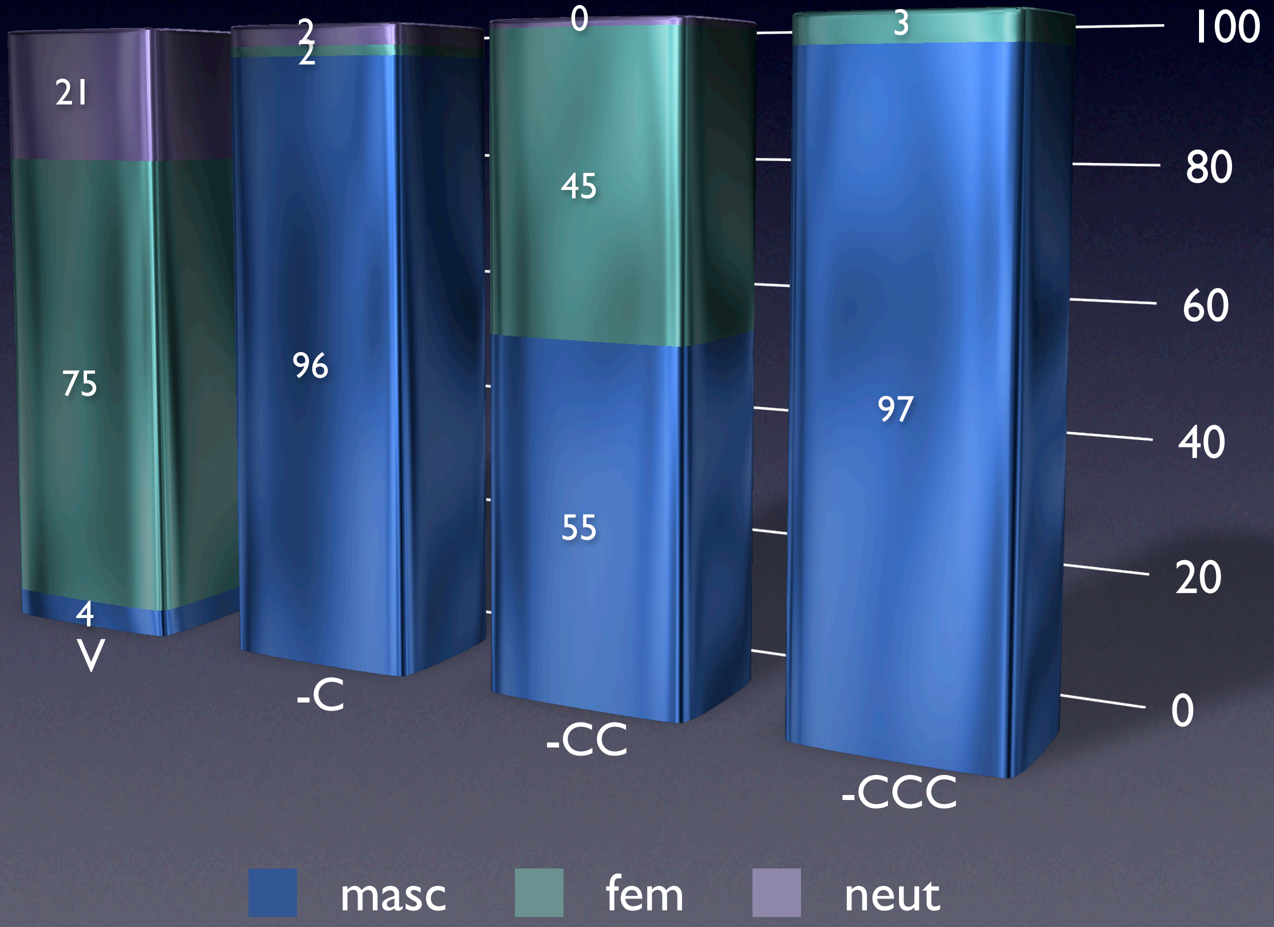
Corpus

- *Uniwersalny Słownik Języka Polskiego* (2006): 200,000 entries, over 42,000 nouns
  - 3 genders: m., f., n. and vacillating
  - 20 declensions: 5 m., 6 f., 6 n. and 3 indeclinable
- types of criteria:
  - sequence length (-C, -CC, -CCC)
  - sequence types (e.g. obstruent + obstruent)
  - specific CC and CCC tokens of types
  - native vs. borrowed vocabulary

Gender and final sequence



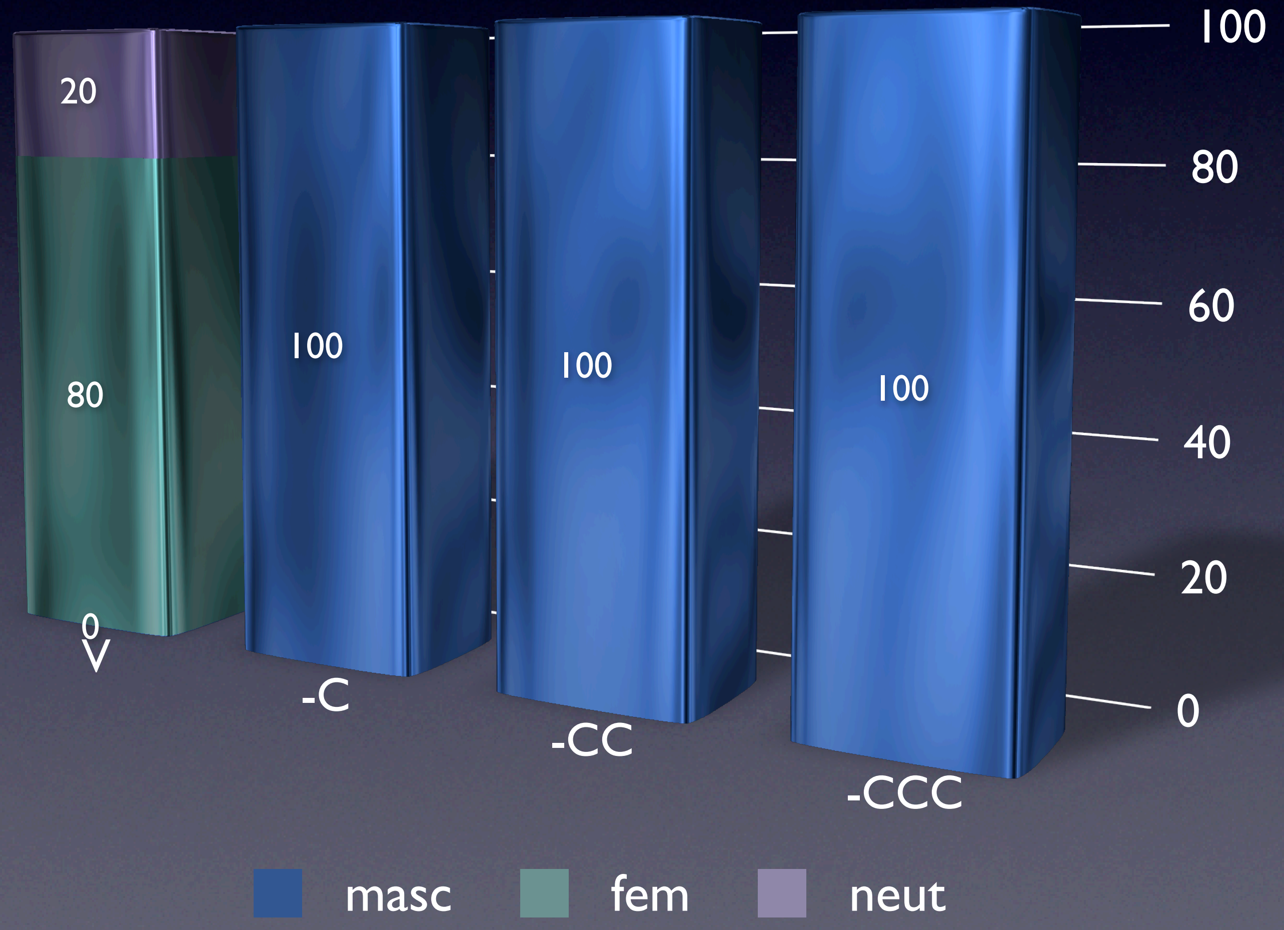
Final sequence and gender



Non-prototypical nouns

- marked phonologically, morphologically and etymologically
  - m. in -V (*poeta* ‘poet’)
  - f. in -(C)(C)C (*biel* ‘white’, *miłość* ‘love’, *pilśń* ‘felt’)
  - n. in -(C)C (*muzeum* ‘museum’, *démarche*)
  - indeclinable nouns (*grizzly, PKB* ‘GNP’)
- m. nouns in -V and f./n. nouns in -(C)(C)C as marked members

Final sequence and gender (prototypical nouns)



Type of final sequence and gender

- soft consonants /ɕ ʑ tɕ dʑ ɲ/ and f. nouns
  - -CC: 98% contain <-ść> (*kość* ‘bone’, *złość* ‘anger’)
  - -CCC: 100% contain 1-2 soft consonants (*pilśń* ‘felt’, *garść* ‘handful’)
- hard consonants and m. nouns
  - fricatives: hard are 20 x more frequent than soft
  - affricates: hard are 4 x more frequent than soft
  - nasals: hard are 10 x more frequent than soft

Final sequence tokens and gender

- m. declensions
  - 152 tokens: 129 doubles and 23 triples
  - c. 25 tokens in most declensions vs. 78 in m4
  - zm (30%), nt (15%), ŋg (4%), st (3%), w̃s (2%)
- f. declensions
  - 19 tokens: 16 doubles and 3 triples
  - ɕtɕ (98%), ɕl (0.3%), ɕɲ / ʑɲ/ ɲtɕ (0.2%)
- 5 most frequent sequences: ɕtɕ, zm, nt, ŋg, st

Native vs. loan sequences

- m. declensions
    - -CC, -CCC esp. in loan nouns
    - m4 (9068): -CC (30%), -CCC (0.5%)
    - -CC in 80% loan tokens (*organizm* ‘organism’, agent, sens ‘sense’)
    - -CCC in 100% loan tokens (*tekst* ‘text’, *punkt* ‘point’)
- f. declensions
    - f5, f6 (2667) – all native (but Rus. *głasnost* ‘glasnost’)

Conclusions

- sequence size: m. as -(C)(C)C vs. f./n. as -V
- sequence type: m. with non-soft C vs. f. with soft C sequences
- tokens: 5 most frequent sequences: /ɕtɕ/ (f.), /zm, nt, ŋg, st/ (m.)
- f. nouns in -(C)(C)C: marked for all analyzed criteria