Quantitative Methods for the Analysis of Classical Japanese Poetry

Kenneth Hanson

April 4, 2014

Corpus Linguistics

- Corpus – a collection of (digitized) text or transcribed speech
  - Useful for linguistic research in a wide variety of fields.
    - Diachronic/historical linguistics
    - Sociolinguistics
    - Pragmatics
  - Useful for quantitative studies.
    - Searching
    - Coding
- Corpus studies of prose are the most common, but what about poetry?

Classical Japanese Poetry

- *Tanka*: the traditional Japanese short poem
  - Structure: five lines with a 5-7-5-7-7 *mora* meter
    - **Mora**: a short syllable, consisting of at most (i) a consonant, (ii) a glide, and (iii) a vowel.
- Case Study: *Ogura Hyakunin Isshu*
  - An anthology of 100 *tanka* from 7th through 13th centuries (Late Old Japanese and Early Middle Japanese) compiled by Fujiwara no Teika.
- Why tanka?
  - Strict structure of the *tanka* format makes it easy to quantify the distribution of syntactic structures.
- Why this text?
  - Representative – a collection of archetypical examples.
  - Source file (from the Japanese Text Initiative) includes line breaks within poems (important later).

秋の田の
かりほの庵の
苫をあらみ
わが衣手は
露にぬれつつ

春過ぎて
夏来にけらし
白妙の
衣ほすてふ
天の香具山

あしびきの
山鳥の尾の
しだり尾の
ながながし夜を
ひとりかもねむ

田子の浦に
打ち出でてみれば
白妙の
富士の高嶺に
雪はふりつつ

A *Tanka* Example

- Hyakunin Isshu Verse 13, by Emperor Yozei

Japanese Text (in Kana)

つくばねの	5
みねよりおつる	7
みなのがわ	5
こいぞつもりて	7
ふちとなりぬる	7

Translation (MacCauley 1917)

From Tsukubane's peak
Falling waters have become
Mina's still, full flow:
So my love has grown to be
Like the river's quiet deeps.

Romanized Text, Gloss, and Literal Translation

Tsukubane-no *mine-yori* *otsuru* *Mina-no-gawa*
Tsukabane-GEN peak-from fall.PRENOM Mina-no-gawa
'The Mina River, which falls from Tsukubane's peak:'

koi-zo *tsumori-te* *fuchi-to* *nari-nuru*
love-EMPH pile.up-CONJ depths-to become-ASP.PRENOM
'my love piles up to become (like) its depths.'

ASP = aspectual auxiliary verb, CONJ = conjunctive particle, EMPH = emphatic particle, GEN = genitive case particle, PRENOM = prenominal form

Goal of the Project

- Some questions we can ask:
  - How does the form of the poetry constrain the syntax?
  - Can the syntax ever override the prescribed form?
- To answer these kinds questions, we need an **annotated corpus**.
  - Part-of-speech (POS) tags are critical.
  - Indexing of line and syllable position would be even better.
- No such corpus exists for Classical Japanese.
  - Solution: build a new annotated corpus.

Building an Annotated Corpus

- Basic method: use a **morphological analyzer** to...
  - Segment the text by *morpheme*, and...
    - **Morpheme**: a word part, either a root, prefix, or suffix
  - Annotate each morpheme with POS, inflection, etc.
- Data source: **Japanese Text Initiative**
  - Provides several poetry anthologies, including Hyakunin Isshu.
  - Files provided are not formatted for computerized analysis.
- Software:
  - Morphological analyzer: **MeCab**
  - Dictionary: **UniDic for Early Middle Japanese** (UniDic-EMJ)

Goals for Corpus Construction

- Things we want to create:
  - Text with spaces inserted
  - Text with spaces and POS tags inserted
  - Table with extensive information on every morpheme
    - Including positional indices

ex. Space-Inserted Text

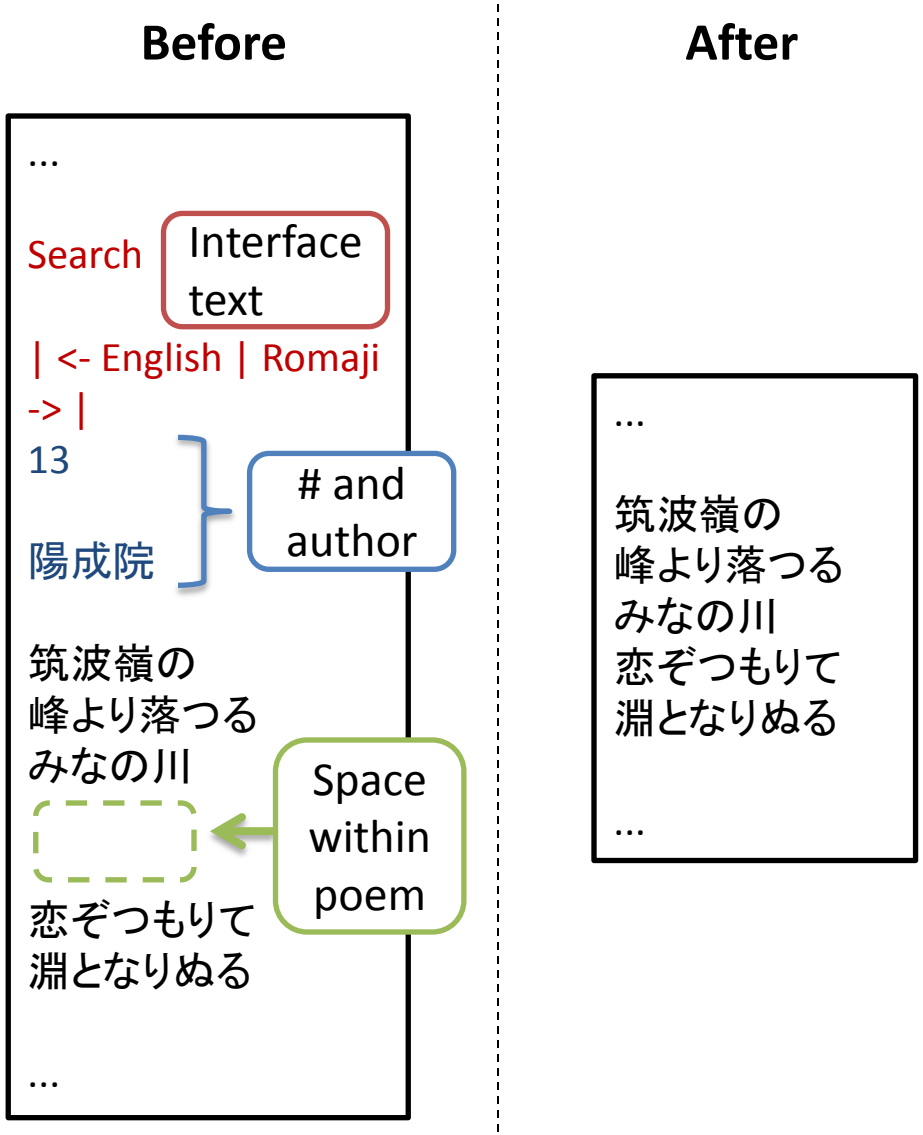
```
Tsukubane no
mine yori otsuru
Mina-no-gawa
koi zo tsumori te
fuchi to nari nuru
```

ex. POS-Tagged Text

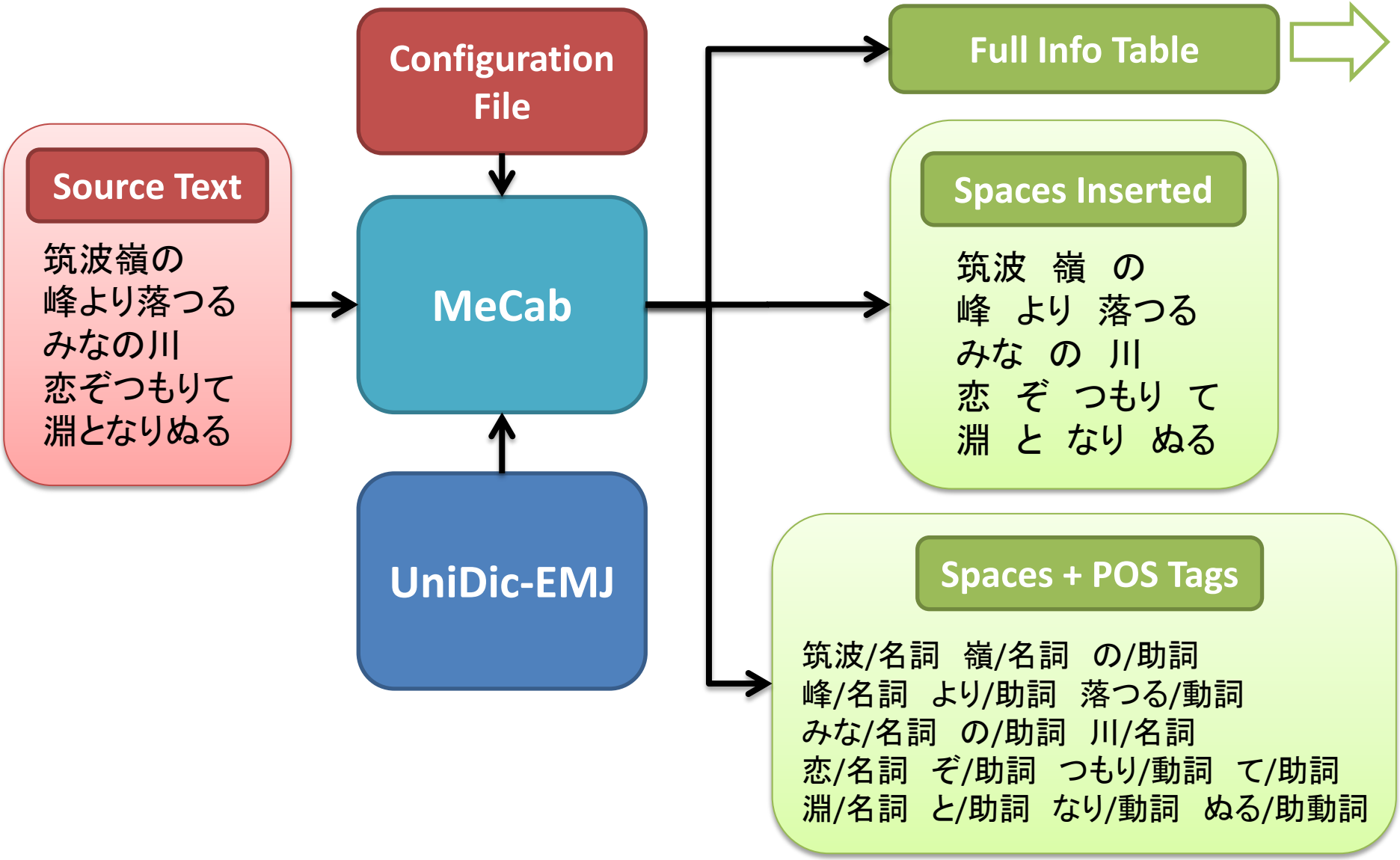
```
Tsukubane/NPR no/P-CASE
mine/N yori/P otsuru/V
Mina-no-gawa/NPR
koi/N zo/P-BND tsumori/V te/P-CONJ
fuchi/N to/P nari/V nuru/AUXV-ASP
```

Preprocessing

- Downloaded source text from the Hyakunin Isshu webpage at University of Virginia Japanese Text Initiative.
- Corrected formatting and removed irrelevant text using regular expressions.



Morphological Analysis



Ex. Full Parse Table from MeCab + UniDic-EMJ

sf	reading	uninfl	lemma	pos1	pos2	pos3	pos4	infl_paradigm	infl_name	origin
筑波	ツクバ	筑波	ツクバ	名詞	固有名詞	地名	一般			固
嶺	ネ	嶺	嶺	名詞	普通名詞	一般				和
の	ノ	の	の	助詞	格助詞					和
EOS										
峰	ミネ	峰	峰	名詞	普通名詞	一般				和
より	ヨリ	より	より	助詞	格助詞					和
落つる	オツル	落つ	落ちる	動詞	一般			文語上二段-タ行	連体形-一般	和
EOS										
みな	ミナ	みな	皆	名詞	普通名詞	副詞可能				和
の	ノ	の	の	助詞	格助詞					和
川	ガワ	川	川	名詞	普通名詞	一般				和
EOS										
恋	コイ	恋	恋	名詞	普通名詞	サ変可能				和
ぞ	ゾ	ぞ	ぞ	助詞	係助詞					和
つもり	ツモリ	つもる	積もる	動詞	一般			文語四段-ラ行	連用形-一般	和
て	テ	て	て	助詞	接続助詞					和
EOS										
淵	フチ	淵	淵	名詞	普通名詞	一般				和
と	ト	と	と	助詞	格助詞					和
なり	ナリ	なる	成る	動詞	非自立可能			文語四段-ラ行	連用形-一般	和
ぬる	ヌル	ぬ	ぬ	助動詞				文語助動詞-ヌ	連体形-一般	和
EOS										
EOS										

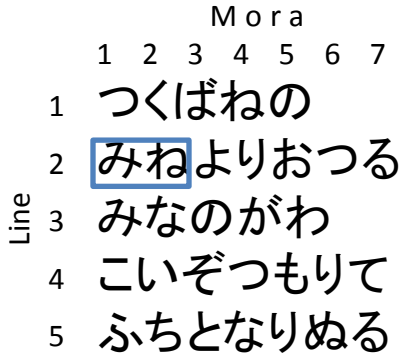
Post-Processing

- Wrote a Python script to extend the full parse results by adding indices to every entry:
  - Poem #
  - Line #
  - Morpheme #
  - Start mora
  - End mora
  - Mora length

Ex. Indexed Parse Table

poem	line	morph	start mora	end mora	length	sf	reading	uninfl	...
13	1	1	1	3	3	筑波	ツクバ	筑波	
13	1	2	4	4	1	嶺	ネ	嶺	
13	1	3	5	5	1	の	ノ	の	
13	2	1	1	2	2	峰	ミネ	峰	
13	2	2	3	4	2	より	ヨリ	より	
13	2	3	5	7	3	落つる	オツル	落つ	
13	3	1	1	2	2	みな	ミナ	みな	
13	3	2	3	3	1	の	ノ	の	
13	3	3	4	5	2	川	ガワ	川	
13	4	1	1	2	2	恋	コイ	恋	
13	4	2	3	3	1	ぞ	ゾ	ぞ	
13	4	3	4	6	3	つもり	ツモリ	つもる	
13	4	4	7	7	1	て	テ	て	
13	5	1	1	2	2	淵	フチ	淵	
13	5	2	3	3	1	と	ト	と	
13	5	3	4	5	2	なり	ナリ	なる	
13	5	4	6	7	2	ぬる	ヌル	ぬ	

Verse 13

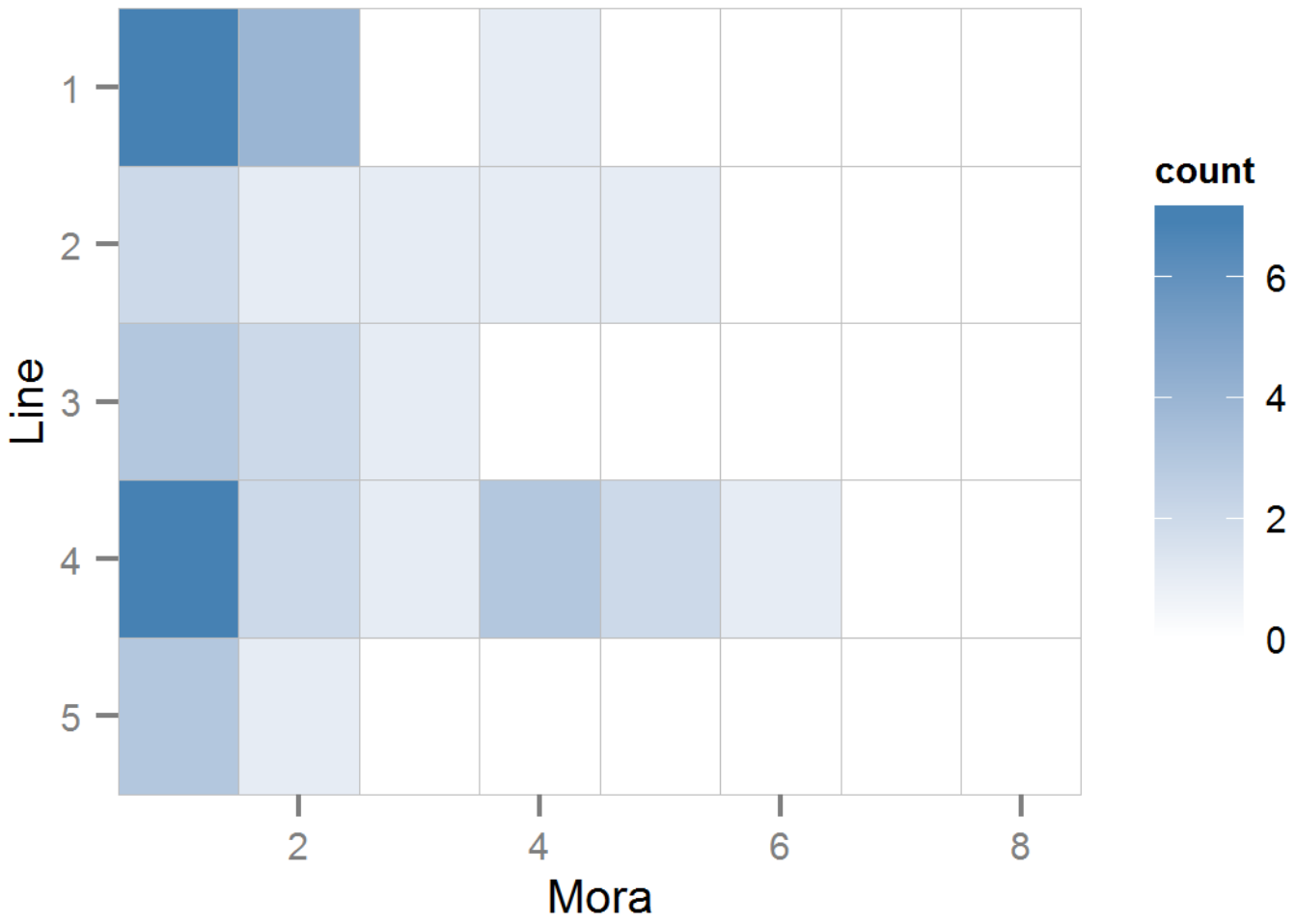


- Wrote Python/R scripts to summarize data by mora and by line.

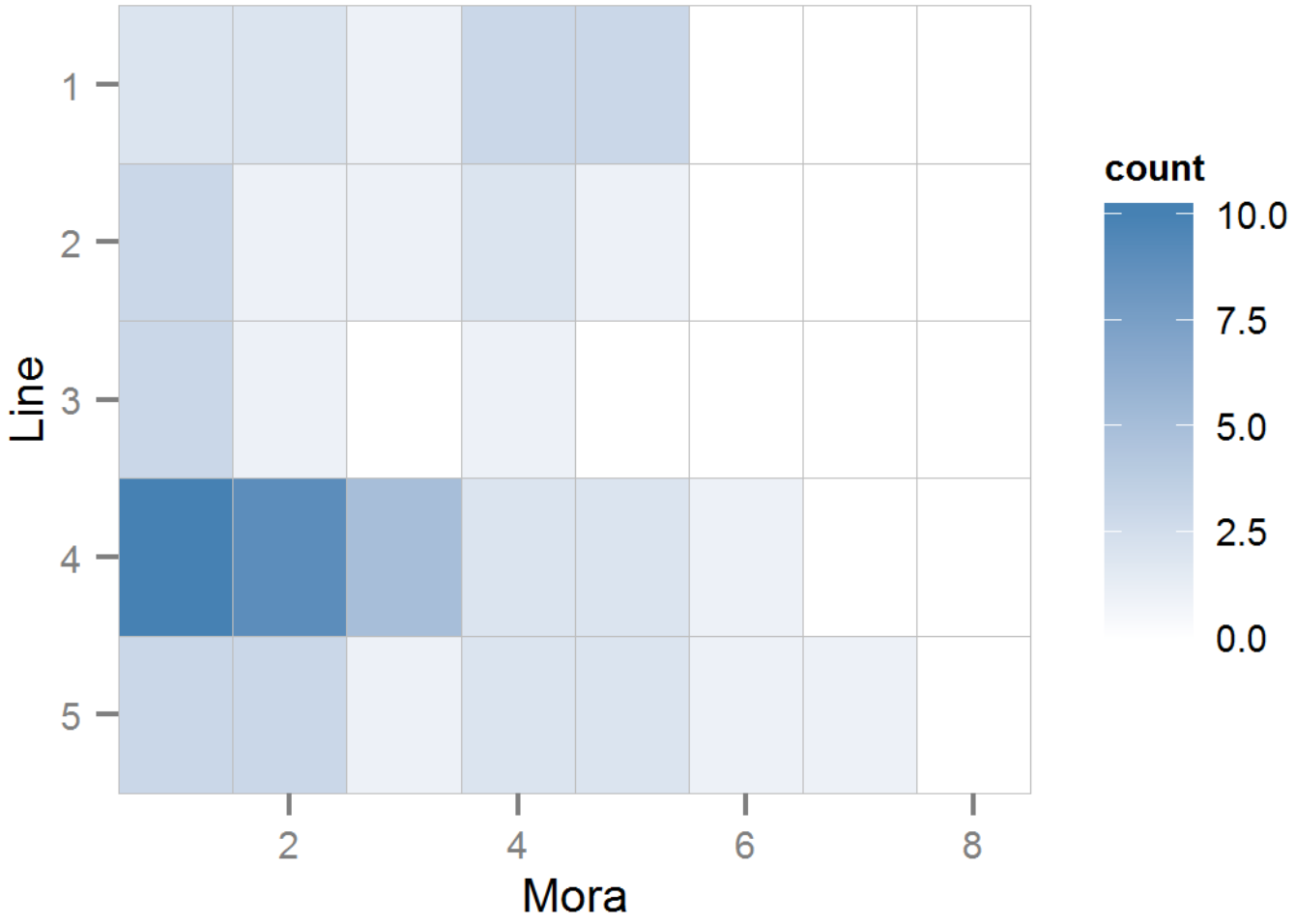
Visualizations and Statistics with R

- One visualization:
  - Heatmaps for the distribution of each part of speech by line and mora
- One statistical experiment:
  - Predicting hypermeter from line contents

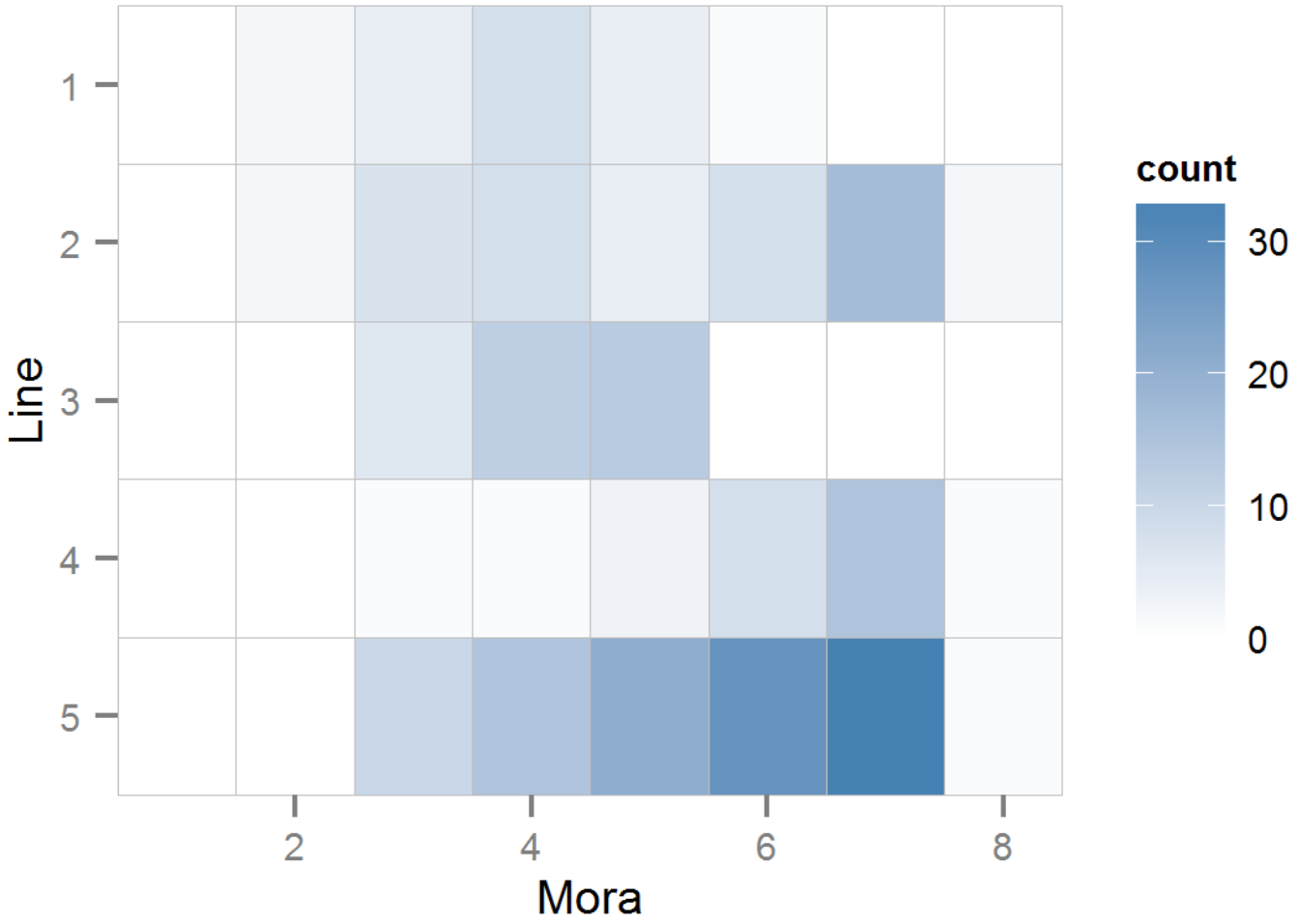
Part-of-Speech Heatmap: Pronouns



Part-of-Speech Heatmap: Adverbs



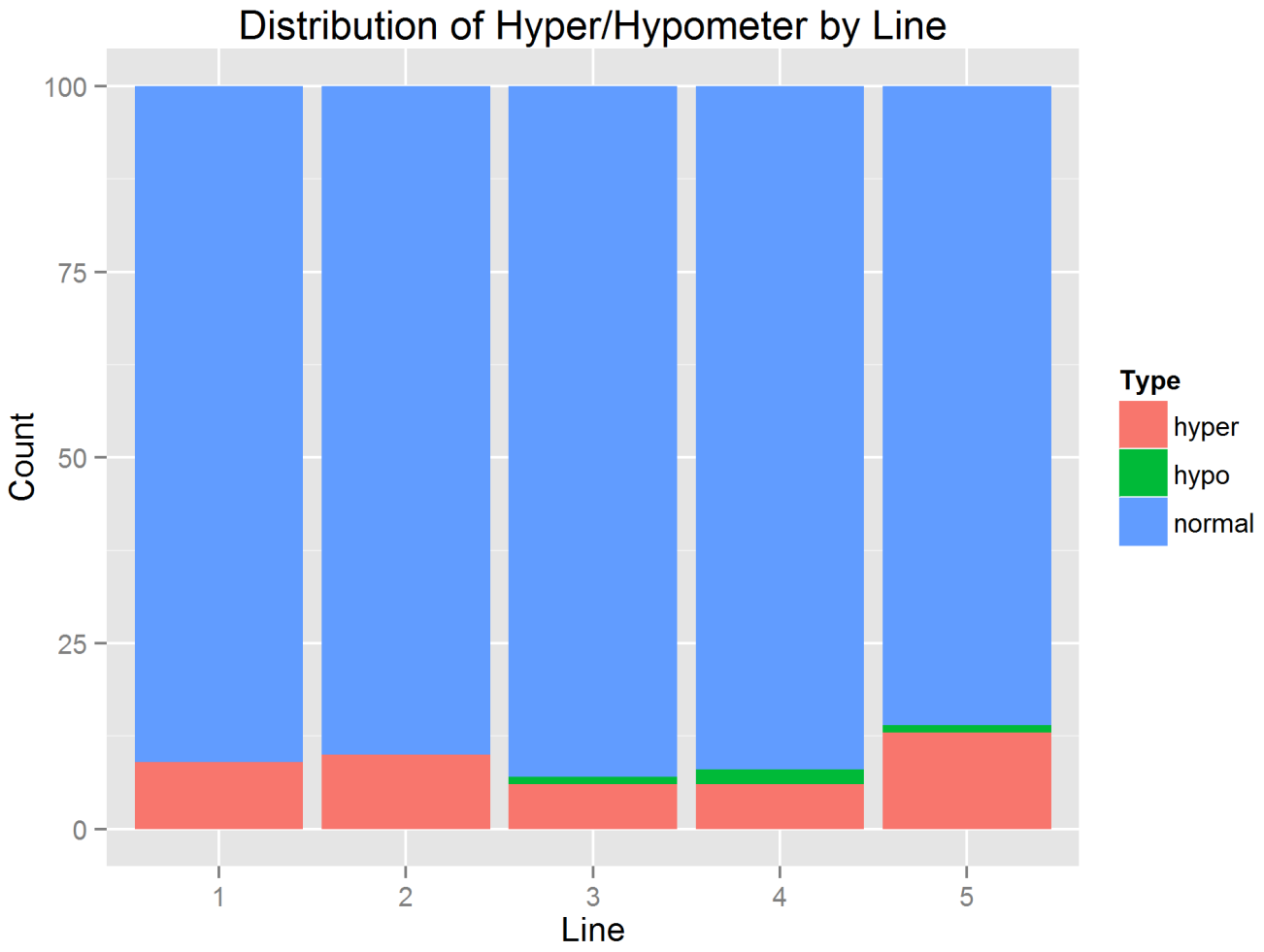
Part-of-Speech Heatmap: Auxiliary Verbs



Predicting Hypermeter

- Hypermeter: one extra mora in a line.
  - Rare in *tanka*, but does occur.
- Conditioning factors to examine:
  - Line number
  - Line contents

Hypermeter and Line number



Hypermeter and Line Contents

In order of decreasing frequency of hypermeter:

Combo	Noun	Pronoun	Adjective	AdjNoun	Verb	AuxVerb	Particle	Adverb	Total	Normal	Hypo	Hyper	hyper.freq
1	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	22	16	0	6	0.2727
2	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	4	3	0	1	0.2500
3	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	64	55	0	9	0.1406
4	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	16	14	0	2	0.1250
5	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	49	43	0	6	0.1224
6	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	139	125	0	14	0.1007
7	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	21	19	0	2	0.0952
8	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	33	30	0	3	0.0909
9	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	17	16	0	1	0.0588
10	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	23	23	0	0	0.0000
11	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	14	13	1	0	0.0000
12	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	12	10	2	0	0.0000
13	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	10	9	1	0	0.0000
14	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	8	8	0	0	0.0000
15	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	7	7	0	0	0.0000
…													

A Statistical Model

```
Call:
glm(formula = type ~ line + Noun + Pronoun + Adjective + AdjNoun +
    Verb + AuxVerb + Particle + Adverb, family = binomial, data = lines.d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7162  -0.4924  -0.3904  -0.1444   2.9997

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.28315    1.12063  -3.822 0.000132 ***
line2          -0.14384    0.50534  -0.285 0.775923
line3          -0.55785    0.56060  -0.995 0.319686
line4          -0.48732    0.56687  -0.860 0.389967
line5           0.24616    0.50563   0.487 0.626373
NounTRUE       -0.29036    0.43357  -0.670 0.503051
PronounTRUE    -0.39695    0.78254  -0.507 0.611973
AdjectiveTRUE  -0.43735    0.65892  -0.664 0.506855
AdjNounTRUE   -14.78415 3042.24628  -0.005 0.996123
VerbTRUE        0.26485    0.40707   0.651 0.515293
AuxVerbTRUE     0.01764    0.41712   0.042 0.966268
ParticleTRUE    2.52483    1.03111   2.449 0.014339 *
AdverbTRUE    -16.07749 1164.56602  -0.014 0.988985
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 297.89  on 499  degrees of freedom
Residual deviance: 272.13  on 487  degrees of freedom
AIC: 298.13

Number of Fisher Scoring iterations: 17
```

Baseline is non-hypermeter

Line number is non-significant

Presence of a particle conditions hypermeter

Summary

- It is possible to create a useful Classical Japanese poetry corpus using currently available software and digitized texts.
- Adding positional indexes to the annotation scheme allows us to analyze the syntax of *tanka* poems.

Acknowledgements

- My thanks to the following people who made this project possible:
  - Catherine Ryu (project advisor and originator of idea)
  - Karthik Durvasula (statistics consulting)

References

- Anthony, L. (2011). AntConc (Version 3.2.2) [Computer Software]. Tokyo: Waseda University. http://www.antlab.sci.waseda.ac.jp/antconc_index.html.
- Clay MacCauley (1917). Hyakunin-Isshu (Single Songs of a Hundred Poets) and Nori no Hatsu-Ne (The Dominant Note of the Law). Yokohama: Kelly and Walsh, Ltd., 1917.
- Japanese Text Initiative. Ogura Hyakunin Isshu. http://etext.lib.virginia.edu/japanese/hyakunin/.
- Kudo, T. (2005). MeCab: Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/.
- Ogiso, T., Komachi, M., Den, Y., & Matsumoto, Y. (2012). UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese. In LREC (pp. 911-915).
- UniDic for Early Middle Japanese. http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%C3%E6%B8%C5%CF%C2%CA%B8UniDic.

Evaluation of Parse Results

- Mostly correct, based on inspection.
- Main problems:
  - Readings for Kanji are sometimes wrong.
  - Proper names are sometimes not detected.
  - Compounds are sometimes analyzed completely rather than left intact as intended.
- All of the above problems are common for human readers, and also when parsing prose, as found by the creators of UniDic-EMJ (Ogiso et al).

sf	reading	uninfl	lemma	pos1	pos2	…
筑波	ツクバ	筑波	ツクバ	名詞	固有名詞	
嶺	ネ	嶺	嶺	名詞	普通名詞	
の	ノ	の	の	助詞	格助詞	
EOS						
峰	ミネ	峰	峰	名詞	普通名詞	
より	ヨリ	より	より	助詞	格助詞	
落つる	オツル	落つ	落ちる	動詞	一般	
EOS						
みな	ミナ	みな	皆	名詞	普通名詞	
の	ノ	の	の	助詞	格助詞	
川	ガワ	川	川	名詞	普通名詞	
EOS						
恋	コイ	恋	恋	名詞	普通名詞	
ぞ	ゾ	ぞ	ぞ	助詞	係助詞	
つもり	ツモリ	つもる	積もる	動詞	一般	
て	テ	て	て	助詞	接続助詞	
EOS						
淵	フチ	淵	淵	名詞	普通名詞	
と	ト	と	と	助詞	格助詞	
なり	ナリ	なる	成る	動詞	非自立可能	
ぬる	ヌル	ぬ	ぬ	助動詞		
EOS						
EOS						