Text Analysis Beyond the Word

Alex Estes & Christopher Hench
UC Berkeley, Department of German

Stakes

Will we ever discover medieval voice?

How do we circumvent irregular orthography and standardization to conduct text analysis on medieval texts?

Why the syllable? Walther von der Vogelweide (E Wa 78)

Frouwe, ir lât iuch nit verdriezen
mîne rede, ob sie gefüege sî.
möht is wider iuch geniezen,
sô wære ich den besten gerne bî.
wizzet, daz ir schœne sît.
habt ir, als ich mich verwæne,
güete bî der wolgetæne?
waz denne an ir einer êren lît!

```
46 word(s)
70 syllable(s)
1.52173913043 syllable(s) per word

17 light syllable(s)
53 heavy syllable(s)
24.2857142857 percent light

34 open syllable(s)
36 closed syllable(s)
48.5714285714 percent open
```

Frauwe ir lat v̊ch nit v̓ dr̊ızzē
mine rede· ob ſie gef̊ůge ſin·
m̊ôht iz wider v̊ch genizzē·
ſo were ich den beſten gerne bi·
wizzet daz ir ſch̊ône ſit·
habt ir als ich mich v̓ wene·
g̊ůte bi der wolgetene·
waz denne an ir ein̓ eren lit·

```
46 word(s)
70 syllable(s)
1.52173913043 syllable(s) per word

25 light syllable(s)
45 heavy syllable(s)
35.7142857143 percent light

31 open syllable(s)
39 closed syllable(s)
44.2857142857 percent open
```

Method

- Syllabification
  - Phonotactics
  - Semiotics and Aesthetics
- Scansion
  - Pedagogy
  - Genre/Author Identification

Syllabification

- Sonority Sequencing Principle (SSP)
- Legality Principle (LP)

Sonority is a property of all sounds
Sonority is highest at the nucleus of a syllable, and lowest at the edges, e.g., ‘trap’:

high sonority a

r

low sonority t p

SSP

tage ‘days’ → 1313 → 13.13 → ta.ge

minne ‘love’ → 23223 → 232.23 → min.ne

entslafen ‘fall asleep’ → 321123132 → 321.123.132 → ent.sla.fen

Tools

SyllabiPy: -uses sonority and user input
-can yield phonotactic frequencies
-works for Latin, in development for more Germanic languages
-can deal with imperfect orthographies

LegaliPy: -requires no user input
-can yield facts about the language
-a good orthography is necessary
- e.g. ‘trap’ → ‘tr’ is a legal onset

Following onsets > .02 percent deemed 'legal':
n - sl - s - br - tr - str - q - ph - w - h - gl - schr - b - g - m - pr - fr - kn - fl - r - p - spr - d - vr - vl - sch - zw - kr - t - sm - sn - c - l - kl - pl - bl - tw - f - v - k - pfl - z - dr - st - sw - cl - sp - j - gr - pf - phl - tj

ist zwî-vel her-zen nâch-ge-bûr daz muoz der sê-le wer-den sûr ge-smæ-het un-de ge-zi-e-ret ist swâ sich par-ri-e-ret un-ver-za-get man-nes muot als a-gel-stern var-we tuot der mac den-noch we-sen geil wand an im sint be-i-diu teil des hi-mels und der hel-le der un-stæ-te ge-sel-le hât die swar-zen var-we gar und wirt och nâch der vin-ster var sô ha-bet sich an die blan-ken der mit stæ-ten ge-dan-ken diz vli-e-gen-de bî-spel ist tum-ben li-u-ten gar ze snel si-ne mu-gens niht er-den-ken wand ez kan vor in wen-ken reh-te al-sam ein schel-lec ha-se zin an-der-halp a-me gla-se ge-lî-chet und des blin-den troum die ge-bent ant-lüt-zes roum doch mac mit stæ-te niht ge-sîn dir-re trü-e-be lîh-te schîn er ma-chet kur-ze frö-u-de al-wâr wer ro-u-fet mich dâ nie kein hâr ge-wuohs in-ne an mî-ner hant der hât vil nâ-he grif-fe er-kant sprich ich gein den vorh-ten och daz glî-chet mî-ner wit-ze doch wil ich tri-we vin-den al-dâ si kan ver-swin-den als viur in dem brun-nen unt daz tou von der sun-nen ouch er-kan-te ich nie sô wî-sen man orn möh-te ger-ne kün-de hân

Following onsets > .02 percent deemed 'legal':
sq - p - str - sc - s - v - fl - gl - h - br - d - t - f - dr - thr - r - n - rh - gr - c - m - g - ch - tr - z - sp - spr - st - ph - cl - q - b - cr - bl - fr - th - phr - pr - l - pl

in no-va fert a-ni-mus mu-ta-tas di-ce-re for-mas cor-po-ra di co-ep-tis nam vos mu-ta-stis et il-las ad-spi-ra-te meis pri-ma-que ab o-ri-gi-ne mun-di ad mea per-pe-tuum de-du-ci-te tem-po-ra car-men an-te ma-re et ter-ras et quod te-git om-nia ca-e-lum u-nus e-rat toto na-tu-rae vul-tus in or-be quem dix-e-re chaos ru-dis in-di-ge-sta-que mo-les nec qu-ic-quam ni-si pon-dus i-ners con-ge-sta-que e-o-dem non be-ne i-unc-ta-rum di-scor-dia se-mi-na re-rum nul-lus ad-huc mun-do pra-e-be-bat lu-mi-na ti-tan nec no-va cre-scen-do re-pa-ra-bat cor-nua pho-e-be nec cir-cum-fu-so pen-de-bat in a-e-re tel-lus pon-de-ri-bus li-bra-ta suis nec brac-chia lon-go mar-gi-ne ter-ra-rum por-rex-e-rat am-phi-tri-te ut-que e-rat et tel-lus il-lic et pon-tus et aer sic e-rat in-sta-bi-lis tel-lus in-na-bi-lis un-da lu-cis e-gens aer nul-

Following onsets > .02 percent deemed 'legal':
dw - fr - pr - kr - þl - g - b - j - sn - q - wl - spr - xr - gr - þw - sl - tr - hv - dr - d - l - swn - wr - sm - phm - n - br - hl - sk - t - z - p - php - þ - w - bl - mk - lk - s - k - sp - sw - r - tw - f - jn - st - mt - hr - hn - th - h - c - þr - m

ak a-na lu-kar-na-sta-þin jah li-u-teiþ al-laim þaim in þam-ma gar-da swa li-uht-jai li-u-haþ iz-war in an-dwa-irþ-ja man-ne ei ga-sa-ihv-a-i-na iz-wa-ra go-da wa-urs-twa jah ha-uh-ja-i-na at-tan iz-wa-ra-na þa-na in hi-mi-nam ni hug-jaiþ ei qem-jau ga-ta-i-ran wi-toþ a-iþ-þau pra-u-fe-tuns ni qam ga-ta-i-ran ak us-full-jan a-men auk qi-þa iz-wis und þa-tei u-sle-i-þiþ hi-mins jah a-ir-þa jo-ta ains a-iþ-þau ains striks ni u-sle-i-þiþ af wi-to-da un-te al-la-ta wa-ir-þiþ iþ saei nu ga-ta-i-riþ a-i-na a-na-bu-sne þi-zo min-ni-sto-no jah la-is-jai swa mans min-ni-sta ha-i-ta-da in þi-u-dan-gard-jai hi-mi-ne iþ saei ta-u-jiþ jah la-is-jai swa sah mi-kils ha-i-ta-da in þi-u-dan-gard-jai hi-mi-ne qi-þa auk iz-wis þa-tei ni-bai ma-na-gi-zo wa-ir-þiþ iz-wa-ra-i-zos ga-ra-ih-teins þau þi-ze bo-kar-je jah fa-re-i-saie ni þau qi-miþ in þi-u-dan-gard-jai hi-mi-ne ba-u-si-de-duþ þa-tei qi-þan ist þaim a-i-ri-zam ni ma-urþr-jais iþ saei ma-ur-

Following onsets > .02 percent deemed 'legal':
cl - m - f - iv - t - j - tr - g - h - ll - i - b - pl - c - p - l - n - ch - fr - cr - fl - s - gl - d - gr - r - br - q - bl - pr - z - v

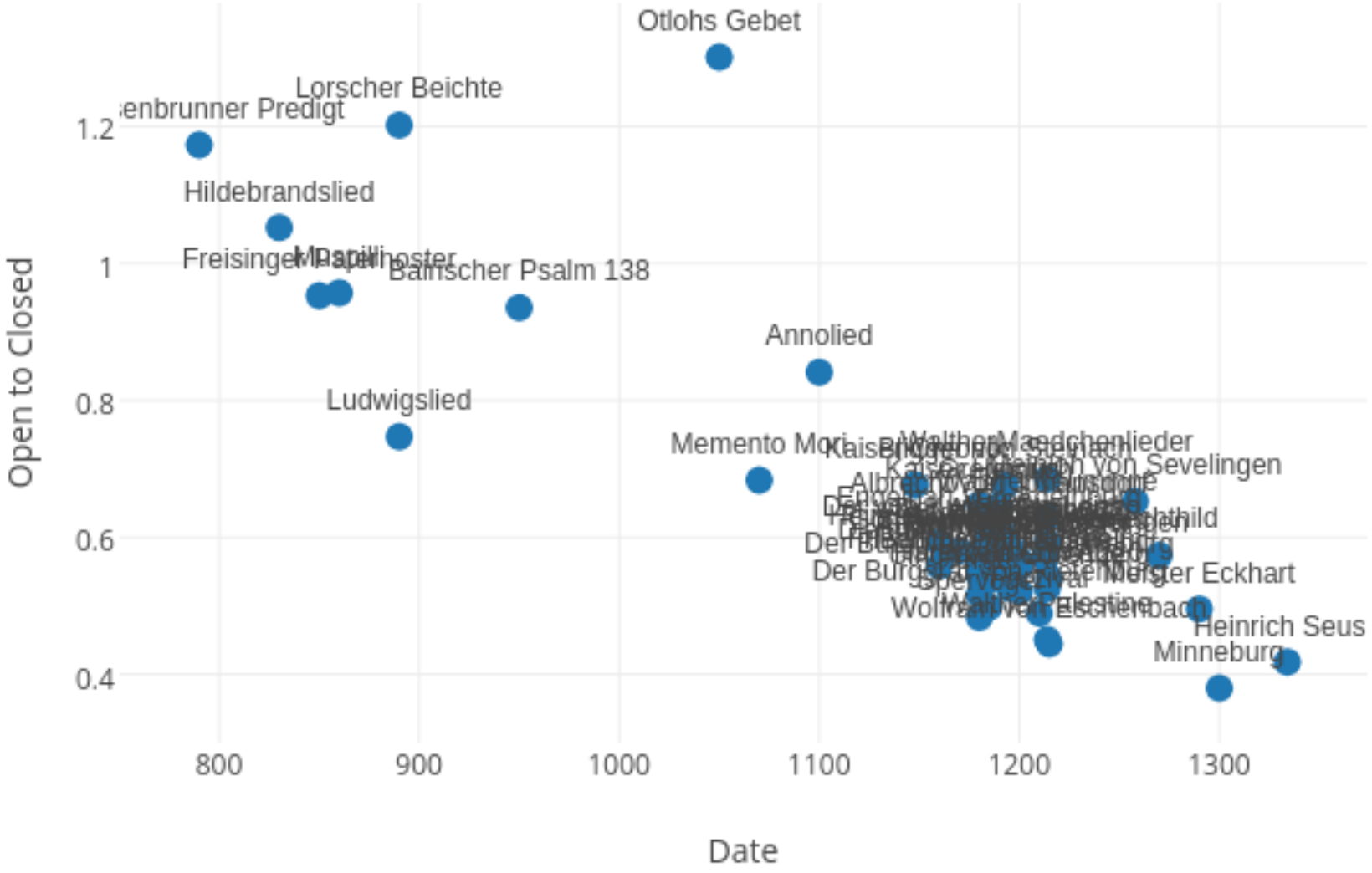
pró-lo-go de-so-cu-pa-do lec-tor sin ju-ra-men-to me pod-rás creer que qu-i-si-e-ra que es-te li-bro co-mo hi-jo del en-ten-di-mi-en-to fu-e-ra el más her-mo-so el más ga-llar-do y más dis-cre-to que pu-di-e-ra i-ma-gi-nar-se pe-ro no he po-di-do yo con-tra-ve-nir al or-den de na-tu-ra-le-za que en e-lla ca-da co-sa en-gend-ra su se-me-jan-te y a-sí qué pod-rá en-gend-rar el es-té-ril y mal cul-ti-va-do in-ge-nio mío si-no la his-to-ria de un hi-jo se-co a-ve-lla-na-do an-to-ja-di-zo y lle-no de pen-sa-mi-en-tos va-rios y nun-ca i-ma-gi-na-dos de o-tro al-gu-no bien co-mo quien se en-gend-ró en u-na cár-cel don-de to-da in-co-mo-di-dad ti-e-ne su a-si-en-to y don-de to-do tris-te ru-i-do ha-ce su ha-bi-ta-ción el so-si-e-go el lu-gar a-pa-ci-ble la a-me-ni-dad de los cam-pos la se-re-ni-dad de los ci-e-los el mur-mu-rar de las fu-en-tes la qu-i-e-tud del es-pí-ri-tu son gran-de par-te pa-ra que las mu-sas más es-té-ri-les se mu-es-tren fe-cun-das y o-frez-can par-tos al mun-do que le col-men de ma-ra-vi-lla y de con-ten-to a-con-te-ce te-ner un pad-re un hi-jo feo y sin gra-cia al-gu-na y el a-mor que le ti-e-ne le po-ne u-na ven-da en los o-jos pa-ra que no vea sus fal-tas an-tes las juz-ga por dis-cre-ci-o-nes y lin-de-zas y las cu-en-ta a sus a-mi-gos por a-gu-de-zas y do-na-i-res pe-ro yo que a-un-que pa-rez-co pad-re soy pad-ras-tro de don qu-i-jo-te no qu-i-e-ro ir-me con la cor-ri-en-te del u-so ni su-pli-car-te ca-si con las lá-gri-mas en los o-jos co-mo o-tros ha-cen lec-tor ca-rí-si-mo que per-do-nes

Following onsets > .02 percent deemed 'legal':
spr - kv - c - lj - sc - vr - tv - r - g - n - sm - sv - skj - k - br - kn - sj - pl - sch - dj - fl - tr - sn - h - skr - b - hskr - dr - nj - kr - str - hm - t - sk - x - j - rdr - sl - gl - gn - tj - gr - q - m - v - d - p - bj - gj - kl - stj - hj - fj - mj - pr - s - f - fr - l - st - bl - sp

för-sta ka-pit-let stock-holm i få-gel-per-spek-tiv det var en af-ton i bör-jan av maj den lil-la träd går-den på mo-se-bac-ke ha-de än-nu ic-ke bli-vit öpp-nad för all-män-he-ten och ra-bat-ter-na vo-ro ej upp-gräv-da snö drop-par-na ha-de ar-be-tat sig upp ge-nom fjo-lå-rets löv sa-mun-gar och höl-lo just på att slu-ta sin kor-ta verk-sam het för att läm-na plats åt de öm-tå-li-ga-re saf-frans blom-mor-na vil-ka ta-git skydd un-der ett o-frukt-samt pä-ron-träd sy-re-ner-na vän-ta-de på syd-lig vind för att få gå i blom men lin-dar-na bjö-do än-nu kär-leks-fil-ter i si-na o-brust-na knop-par åt bo-fin-kar-na som bör-jat byg-ga si-na lav-kläd-da bon mel-lan stam och gren än-nu ha-de in-gen män-sko-fot tram-pat sand-gån-gar-na se-dan si-sta vin-terns snö gått bort och där-för lev-des ett o-be-svä-rat liv dä-rin-ne av

Following onsets > .02 percent deemed 'legal':
ch - ph - thr - g - gl - l - cr - n - s - spr - dw - sq - th - t - tw - st - sc - tr - mr - scr - v - pl - j - kn - sn - q - c - dr - sw - gr - fl - b - w - wr - p - br - cl - r - fr - h - spl - k - pr - bl - m - f - sp - sh - wh - sk - sm - sch - str - d - sl

my fa-ther had a small e-sta-te in not-ting-ham-shi-re i was the third of fi-ve sons he sent me to em-ma-nuel col-le-ge in cam-brid-ge at fo-ur-teen years old whe-re i re-si-ded three years and ap-plied my-self clo-se to my stu-dies but the char-ge of ma-in-ta-i-ning me al-though i had a ve-ry scan-ty al-lo-wan-ce being too great for a nar-row for-tu-ne i was bound ap-pren-ti-ce to ja-mes ba-tes an e-mi-nent sur-geon in lon-don with whom i con-ti-nued four years and my fa-ther now and then sen-ding me small sums of mo-ney i laid them out in le-ar-ning na-vi-ga-tion and o-ther parts of the ma-the-ma-tics u-se-ful to tho-se who in-tend to tra-vel as i al-ways be-li-e-ved it would be so-me ti-me or o-ther


Otlohs Gebet
Lorscher Beichte
enbrunner Predigt
Hildebrandslied
Freisinger Paternoster
Bairischer Psalm 138
Annolied
Ludwigslied
Memento Mori
Walther
Maedchenlieder
Meister Eckhart
Heinrich Seus
Minneburg
Open to Closed
Date
1.2
1
0.8
0.6
0.4
800
900
1000
1100
1200
1300

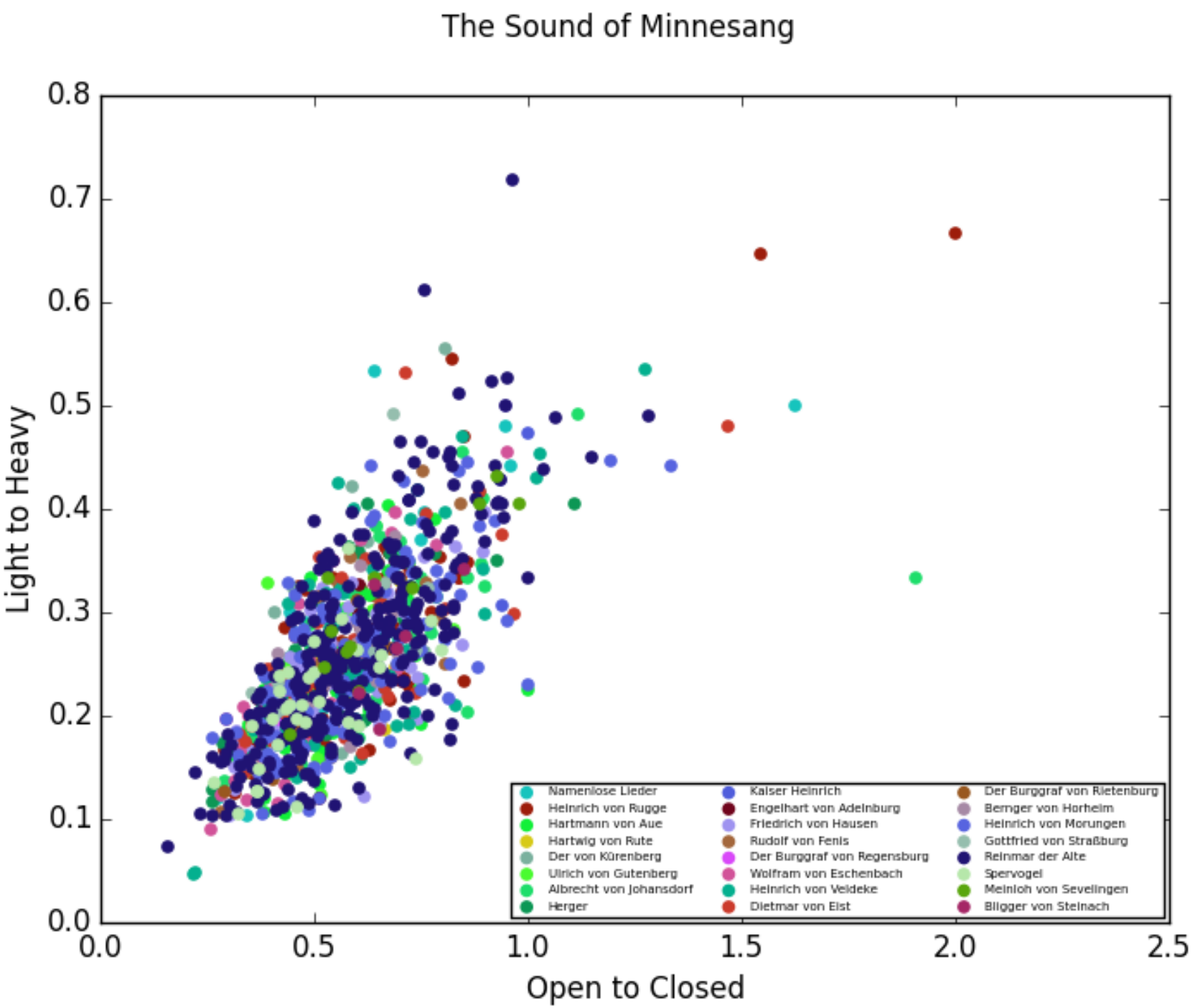
Pater noster

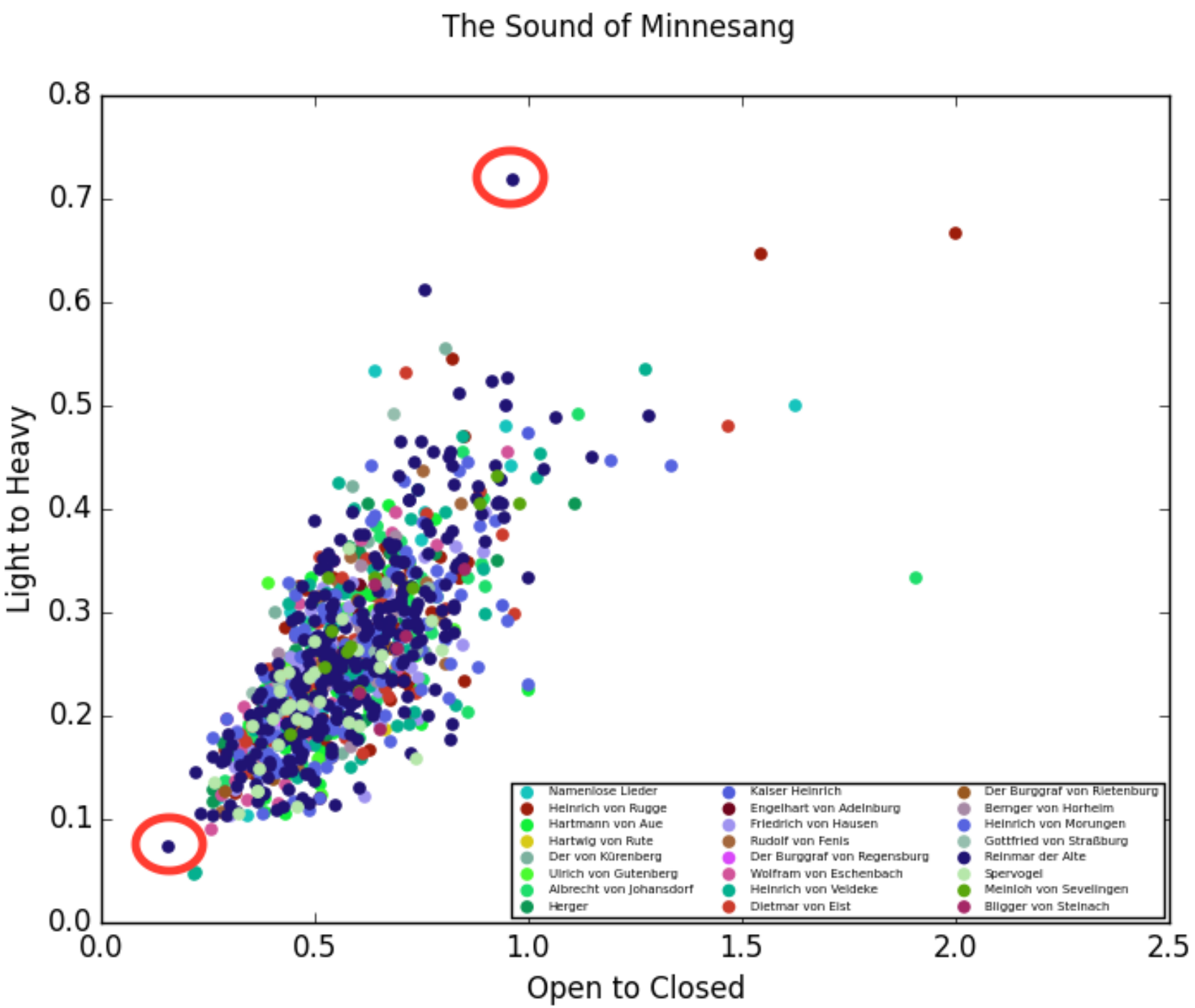
9th Century OHG:

Fater unser, thu in himilom bist, giuuihit si namo thin. Quaeme richi thin, uuerdhe uuilleo thin, sama so in himile endi in erthu, Broot unseraz emezzigaz gib uns hiutu, endi farlaz uns sculdhi unsero, sama so uuir farlazzem scolom unserem endi ni gileidi unsih in costunga, auh arlösi unsih fona ubile.

13th Century MHG:

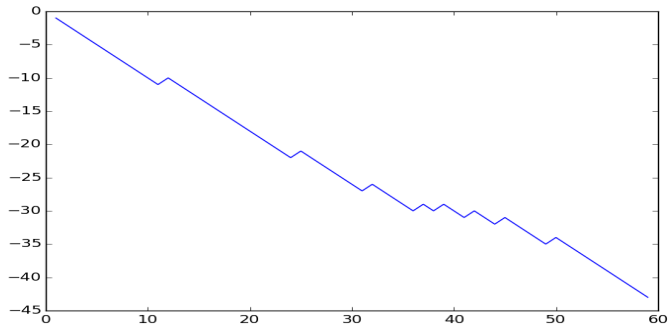
Vater unser, der du in dem himel bist, geheileget si din nam [...], zuo kum an uns daz riche din, din wille werde hie als in dim riche. Din götlich bröt daz gip uns hiute [...], vergip uns unser schult, also wir unsern schuldern han, behorunge uns laz anic sin loese uns von disen übeln


The Sound of Minnesang
Light to Heavy
Open to Closed
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.5
1.0
1.5
2.0
2.5
Namenlose Lieder
Heinrich von Rugge
Hartmann von Aue
Hartwig von Rute
Der von Kürenberg
Ulrich von Gutenberg
Albrecht von Johansdorf
Herger
Kaiser Heinrich
Engelhart von Adelnburg
Friedrich von Hausen
Rudolf von Fenis
Der Burggraf von Regensburg
Wolfram von Eschenbach
Heinrich von Veldeke
Dietmar von Elst
Der Burggraf von Rietenburg
Berngер von Horheim
Heinrich von Morungen
Gottfried von Straßburg
Reinmar der Alte
Spervogel
Meinloh von Sevelingen
Bligger von Steinach


The Sound of Minnesang
Light to Heavy
Open to Closed
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.5
1.0
1.5
2.0
2.5
Namenlose Lieder
Heinrich von Rugge
Hartmann von Aue
Hartwig von Rute
Der von Kürenberg
Ulrich von Gutenberg
Albrecht von Johansdorf
Herger
Kaiser Heinrich
Engelhart von Adelnburg
Friedrich von Hausen
Rudolf von Fenis
Der Burggraf von Regensburg
Wolfram von Eschenbach
Heinrich von Veldeke
Dietmar von Eist
Der Burggraf von Rietenburg
Berngер von Horheim
Heinrich von Morungen
Gottfried von Straßburg
Reinmar der Alte
Spervogel
Meinloh von Sevelingen
Bligger von Steinach

Reinmar der Alte (192,25-214C)

Dêst ein nôt, daz mich ein man
vor al der werlte twinget, swes er wil
sol ich, des ich niht enkan,
beginnen, daz ist mir ein swaerez spil
Ich hât ie vil staeten muot
nu muoz ich leben als ein wîp,
diu minnet und daz angestlîchen tuot



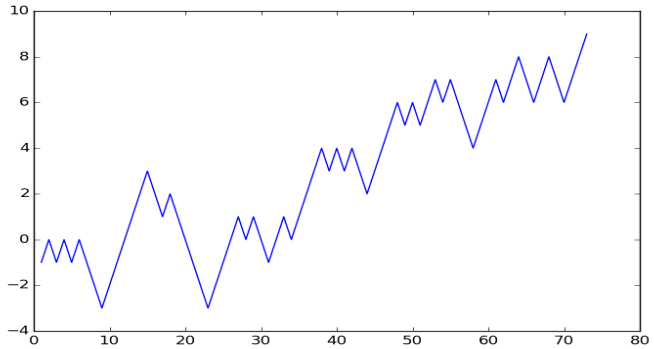
47 word(s)
59 syllable(s)
1.25531914894 syllable(s) per word

4 light syllable(s)
55 heavy syllable(s)
6.77966101695 percent light

9 open syllable(s)
50 closed syllable(s)
15.2542372881 percent open

Reinmar der Alte (173,13-96C,53b)

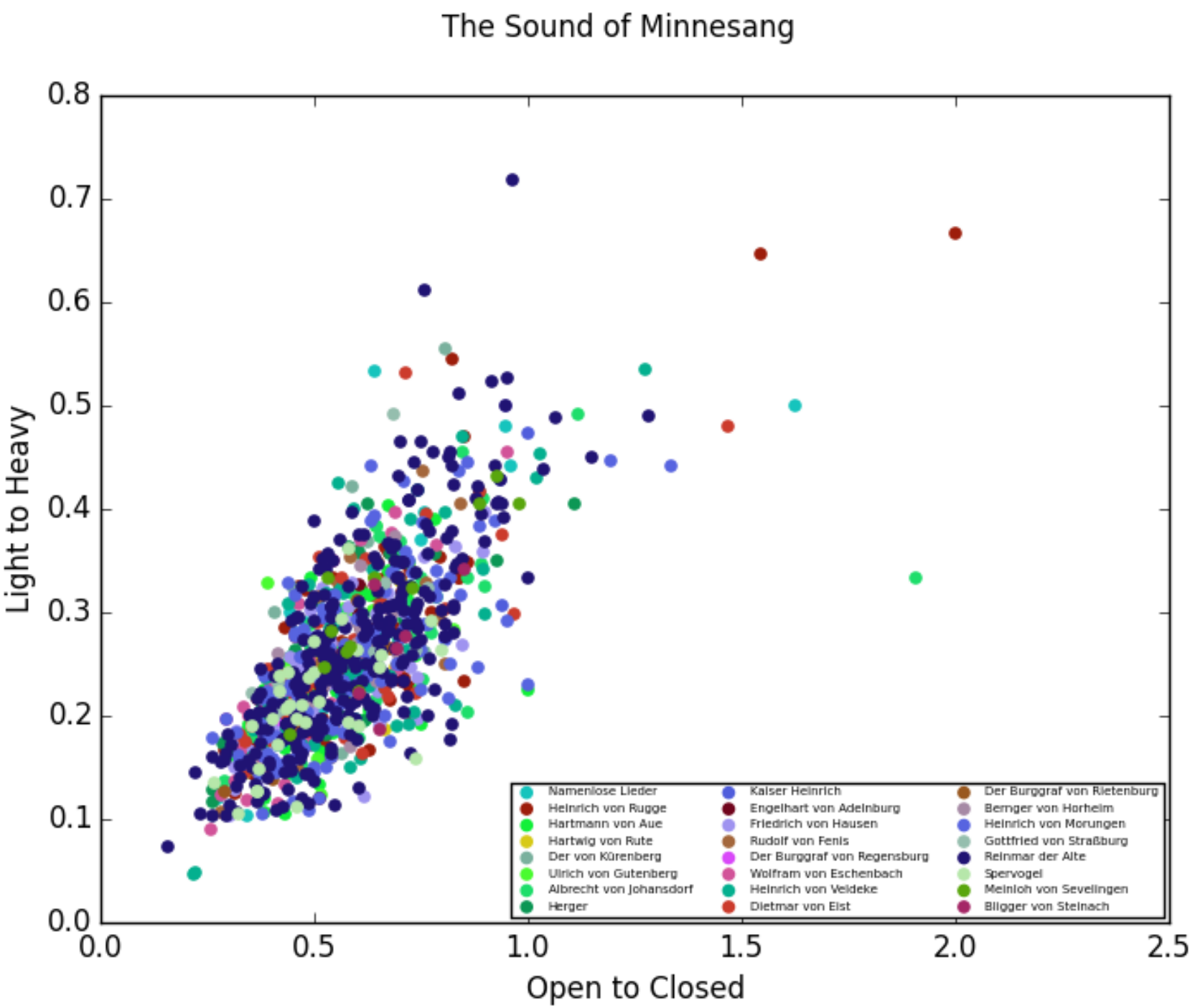
Swenne ich sî mit mîner valschen rede betrüge,
sô het ich sî unreht erkant
und gevâhe sî mich iemer an deheiner lüge,
sâ sô schupfe mich zehant
Und geloube niemer mîner klage,
dar zuo niht, des ich sage
dâ vor müeze mich got behüeten alle tage

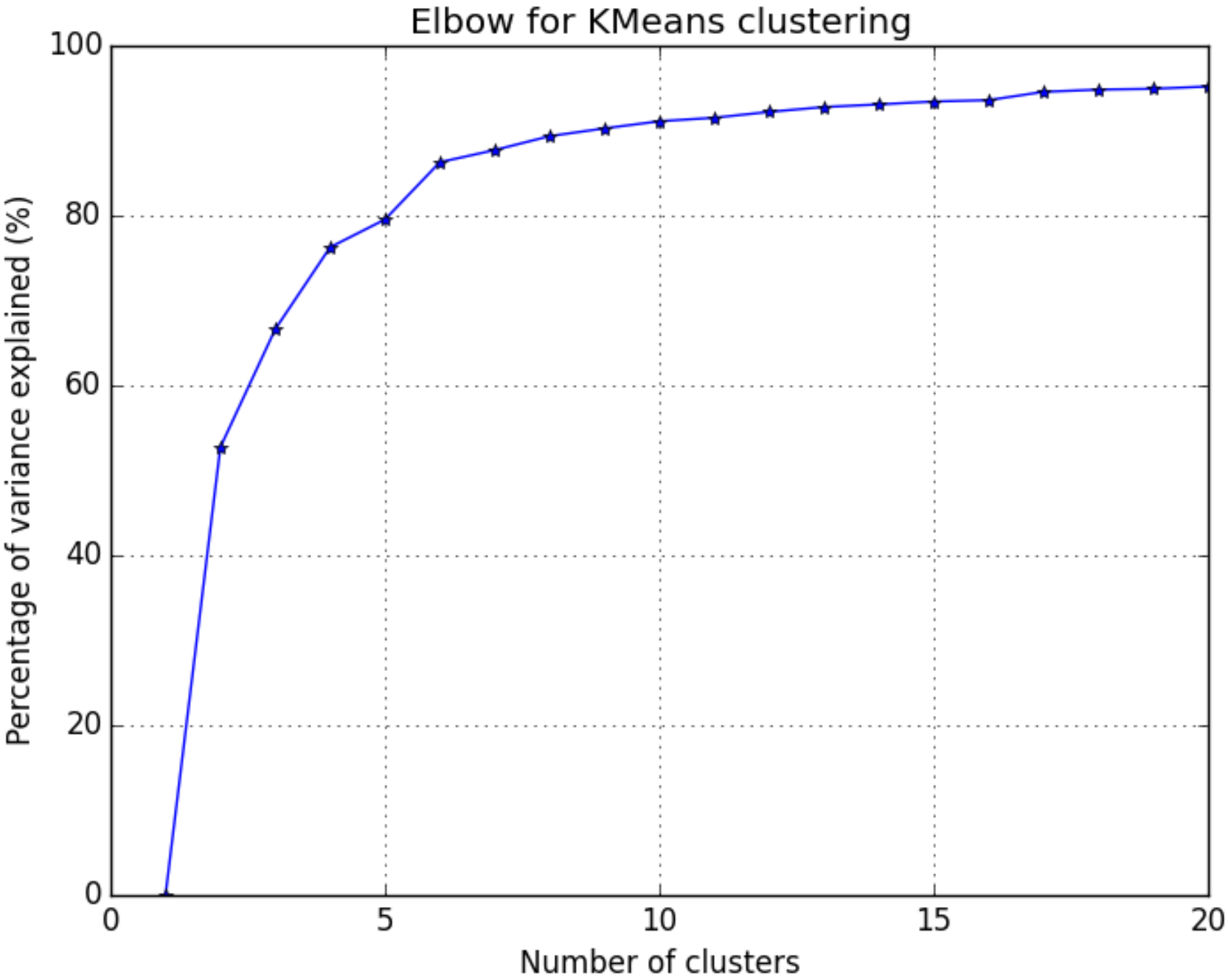


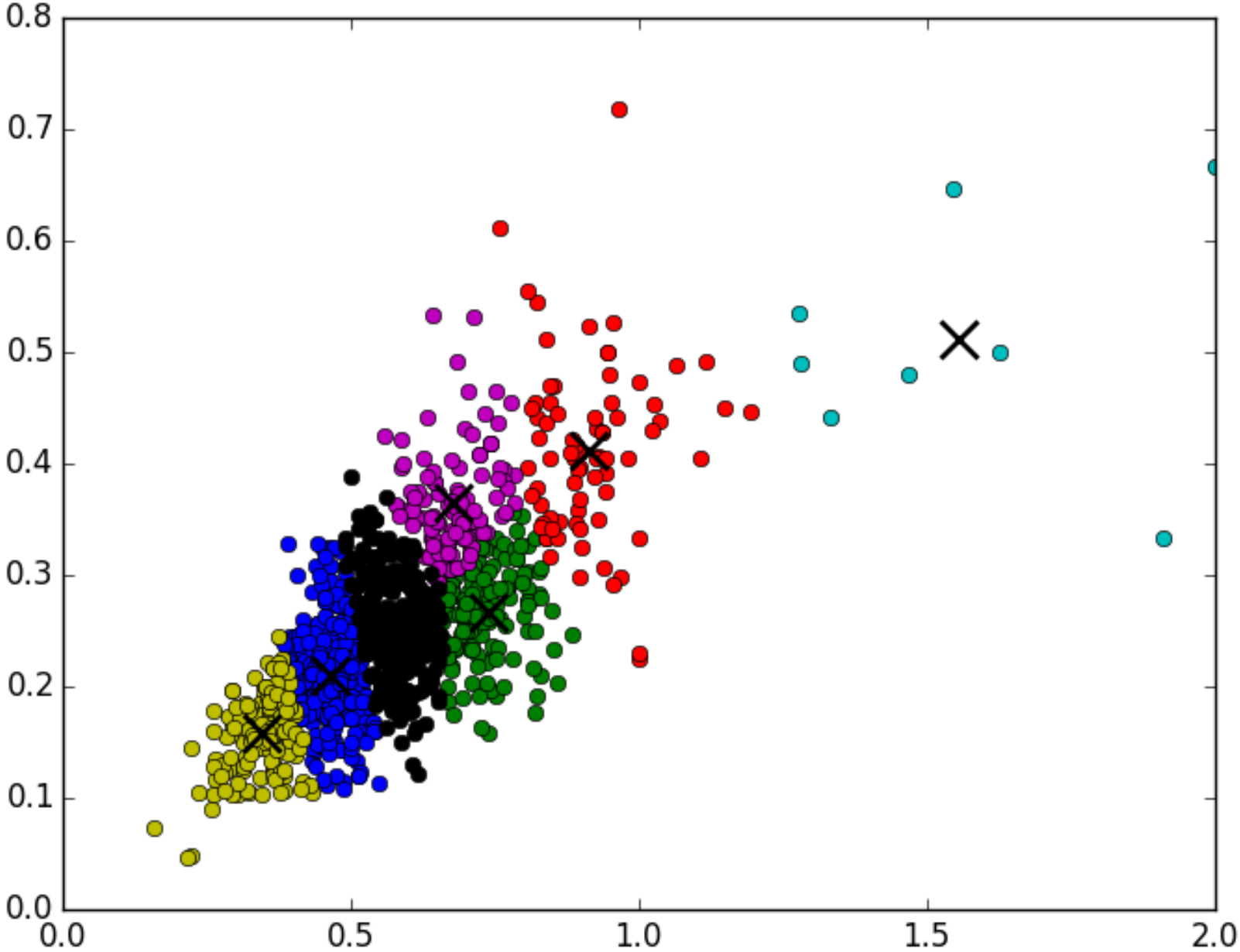
46 word(s)
73 syllable(s)
1.58695652174 syllable(s) per word

25 light syllable(s)
48 heavy syllable(s)
34.2465753425 percent light

42 open syllable(s)
31 closed syllable(s)
57.5342465753 percent open


The Sound of Minnesang
Light to Heavy
Open to Closed
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.5
1.0
1.5
2.0
2.5
Namenlose Lieder
Heinrich von Rugge
Hartmann von Aue
Hartwig von Rute
Der von Kürenberg
Ulrich von Gutenberg
Albrecht von Johansdorf
Herger
Kaiser Heinrich
Engelhart von Adelnburg
Friedrich von Hausen
Rudolf von Fenis
Der Burggraf von Regensburg
Wolfram von Eschenbach
Heinrich von Veldeke
Dietmar von Elst
Der Burggraf von Rietenburg
Berngер von Horheim
Heinrich von Morungen
Gottfried von Straßburg
Reinmar der Alte
Spervogel
Meinloh von Sevelingen
Bligger von Steinach


Elbow for KMeans clustering
Percentage of variance explained (%)
Number of clusters
0
20
40
60
80
100
0
5
10
15
20


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
2.0

Scansion

- 4 trochees (stressed-unstressed syllable pair) per line
- sometimes a syllable takes the place of two metrical syllables
- sometimes two syllables take the place of one metrical syllables
- there are also restrictions on syllable quality in certain positions
- anacrusis allows for extrametrical syllables
- there is no algorithm for this

Tagging

BEGL/BEGL WBY/WBY ein/MORA WBY/WBY rî/MORA_HAUPT ter/MORA WBY/WBY sô/MORA_HAUPT WBY/WBY ge/MORA lê/MORA_HAUPT ret/MORA WBY/WBY was/MORA_HAUPT WBY/WBY ENDL/ENDL

BEGL/BEGL WBY/WBY daz/MORA_HAUPT WBY/WBY er/MORA WBY/WBY an/MORA_HAUPT WBY/WBY den/MORA WBY/WBY buo/MORA_HAUPT chen/MORA WBY/WBY las/MORA_HAUPT WBY/WBY ENDL/ENDL

BEGL/BEGL WBY/WBY swaz/MORA WBY/WBY er/MORA_HAUPT WBY/WBY dar/MORA WBY/WBY an/MORA_HAUPT WBY/WBY ge/MORA schri/MORA_HAUPT ben/MORA WBY/WBY vant/MORA_HAUPT WBY/WBY ENDL/ENDL

BEGL/BEGL WBY/WBY der/MORA_HAUPT WBY/WBY was/MORA WBY/WBY hart/DOPPEL man/MORA_NEBEN WBY/WBY ge/MORA nant/MORA_HAUPT WBY/WBY ENDL/ENDL

BEGL/BEGL WBY/WBY dienst/MORA_HAUPT man/MORA WBY/WBY was/MORA_HAUPT WBY/WBY er/MORA WBY/WBY zuo/EL ou/DOPPEL we/MORA_NEBEN WBY/WBY ENDL/ENDL

BEGL/BEGL WBY/WBY der/MORA WBY/WBY nam/MORA_HAUPT WBY/WBY im/MORA WBY/WBY ma/HALB_HAUPT ne/HALB ge/MORA WBY/WBY schou/DOPPEL we/MORA_NEBEN WBY/WBY ENDL/ENDL

Scansion Output (Parzival ll. 502-506)

502: von-a-râ-bîe-des-gol-des
X / X' X / X` X / ---' / X` Zweisilbig weiblich

503: he-ter-ma-ne-gen-knol-len-brâht
/ X' X / X' ◡ ◡ / X' X / X' Einsilbig männlich

504: liu-te-vin-ster-sô-diu-naht
/ X' X / X' X / X' X / X' Einsilbig männlich

505: wârn-al-le-die-von-za-za-manc
X / X' X / X' X / X' X / X` Einsilbig männlich

506: bî-den-dûht-in-diu-wî-le-lanc
/ X' X / X' ◡ ◡ / X' X / X' Einsilbig männlich

Further Uses?

- Syllables
- Meter