A Rule-Based Parsing System for Encoding Kennicott's Collation of the Hebrew Bible

Luigi Bambaci[1]

[1]University of Bologna
Department of Cultural Heritage

12 January 2020

Plan

- Introduction
  - Qohelet
  - Kennicott's work
- Digitalization
  - Optical character recognition
  - The Context-free Grammar
  - The Visitor
  - TEI encoding
- Stemmatic analysis

Qohelet: A Digital Scholarly Edition

The parser is part of a PhD project devoted to the preparation of a digital scholarly edition of the biblical book of Qohelet

Goals of the edition

- *Collatio*: digitalization of readings
- *Constitutio textus*: critical text (eclectic edition)
- Encoding according to the Text Encoding Initiative

https://tei-c.org/

Qohelet: A Digital Scholarly Edition

Goals of the parser

- The parser is designed to encode the readings of Qohelet contained in medieval mss and printed editions as collated by Kennicott
- The encoding is used for quantitative analysis
- To identify stemmatic relationships

Qohelet: A Digital Scholarly Edition

Goals of the parser

- The parser is designed to encode the readings of Qohelet contained in medieval mss and printed editions as collated by Kennicott
- The encoding is used for quantitative analysis
- To identify stemmatic relationships

http://cophilab.ilc.cnr.it/

http://www.ilc.cnr.it/

The book of Qohelet

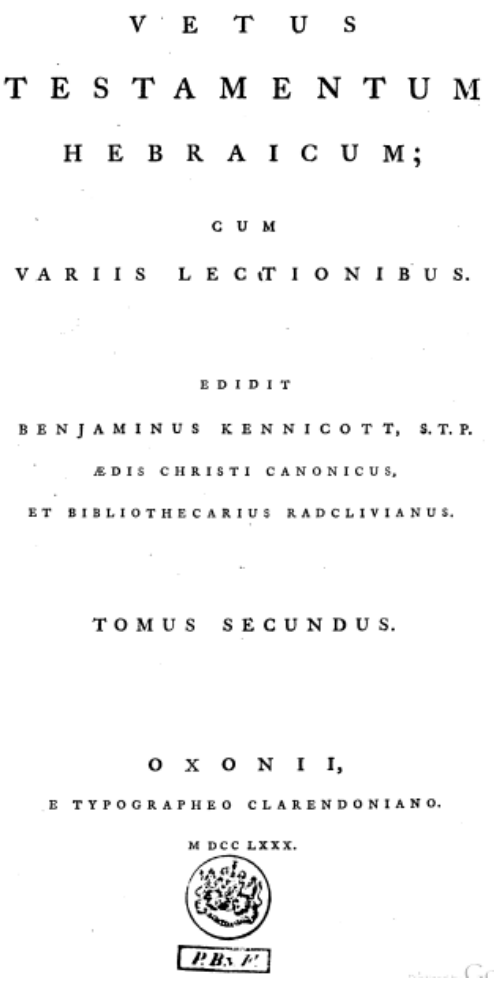
The book of Qohelet or Ecclesiastes is one of the books of the Hebrew Bible (V-III a.e.v.)

Sources

1. Direct sources: witnesses in Hebrew
   - Qumran scrolls (II-I a.e.v.)
   - Hebrew medieval manuscripts and printed editions
2. Indirect sources: ancient translations (Versions):
   - 2.1 **Primary translations** (Hebrew source):
     - Greek (Septuaginta, I-II e.v.)
     - Syriac (Peshitta, II e.v.)
     - Latin (Vulgata, *Commentarius*, IV e.v.)
     - Aramaic (Targumim, VII e.v.)
   - 2.2 **Secondary translations** (Greek source):
     - Ethiopic (IV e.v.)
     - Armenian (V e.v.)
     - Syro-hexapla (VII e.v.)
     - ...

Medieval tradition: the XVIII c. collations

Our access to medieval documentation is still mainly mediated by the works of B. Kennicott and G. B. De Rossi



V E T U S

T E S T A M E N T U M

H E B R A I C U M;

C U M

V A R I I S L E C T I O N I B U S.

E D I D I T

B E N J A M I N U S K E N N I C O T T, S. T. P.

Æ D I S C H R I S T I C A N O N I C U S,

E T B I B L I O T H E C A R I U S R A D C L I V I A N U S.

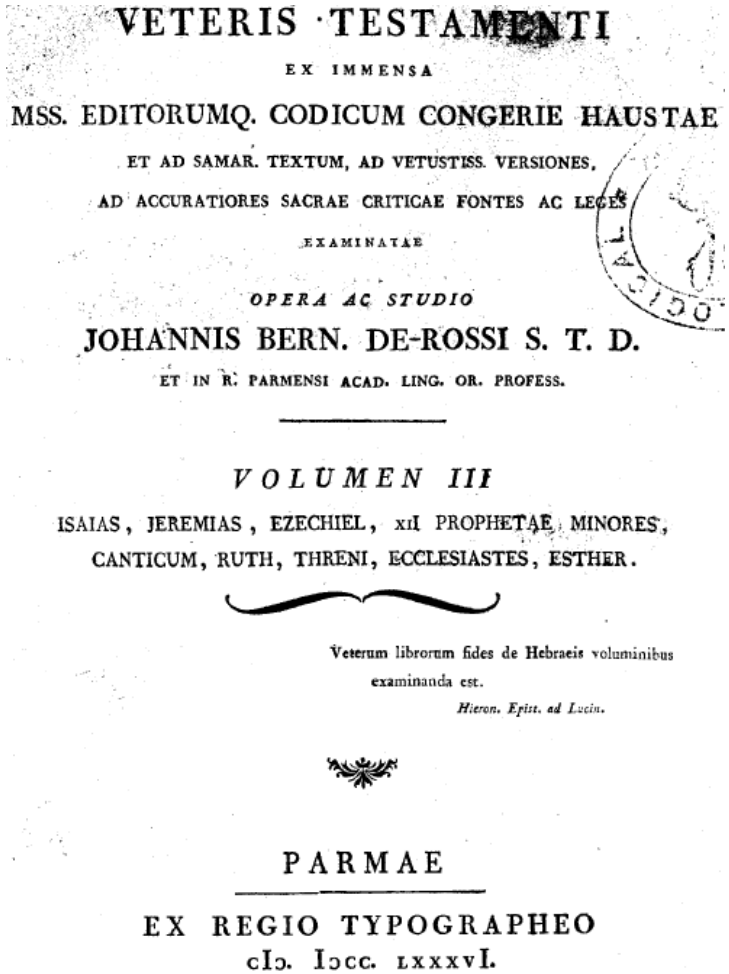
T O M U S S E C U N D U S.

O X O N I I,

E T Y P O G R A P H E O C L A R E N D O N I A N O.

M DCC LXXX.

Kennicott, 1778



VETERIS TESTAMENTI

EX IMMENSA

MSS. EDITORUMQ. CODICUM CONGERIE HAUSTAE

ET AD SAMAR. TEXTUM, AD VETUSTISS. VERSIONES,

AD ACCURATIORES SACRAE CRITICAE FONTES AC LEGES

EXAMINATAE

OPERA AC STUDIO

JOHANNIS BERN. DE-ROSSI S. T. D.

ET IN R. PARMENSI ACAD. LING. OR. PROFESS.

VOLUMEN III

ISAIAS, JEREMIAS, EZECHIEL, XII PROPHETAE MINORES,

CANTICUM, RUTH, THRENI, ECCLESIASTES, ESTHER.

Veterum librorum fides de Hebraeis voluminibus

examinanda est.

Hieron. Epist. ad Lucin.

PARMAE

EX REGIO TYPOGRAPHEO

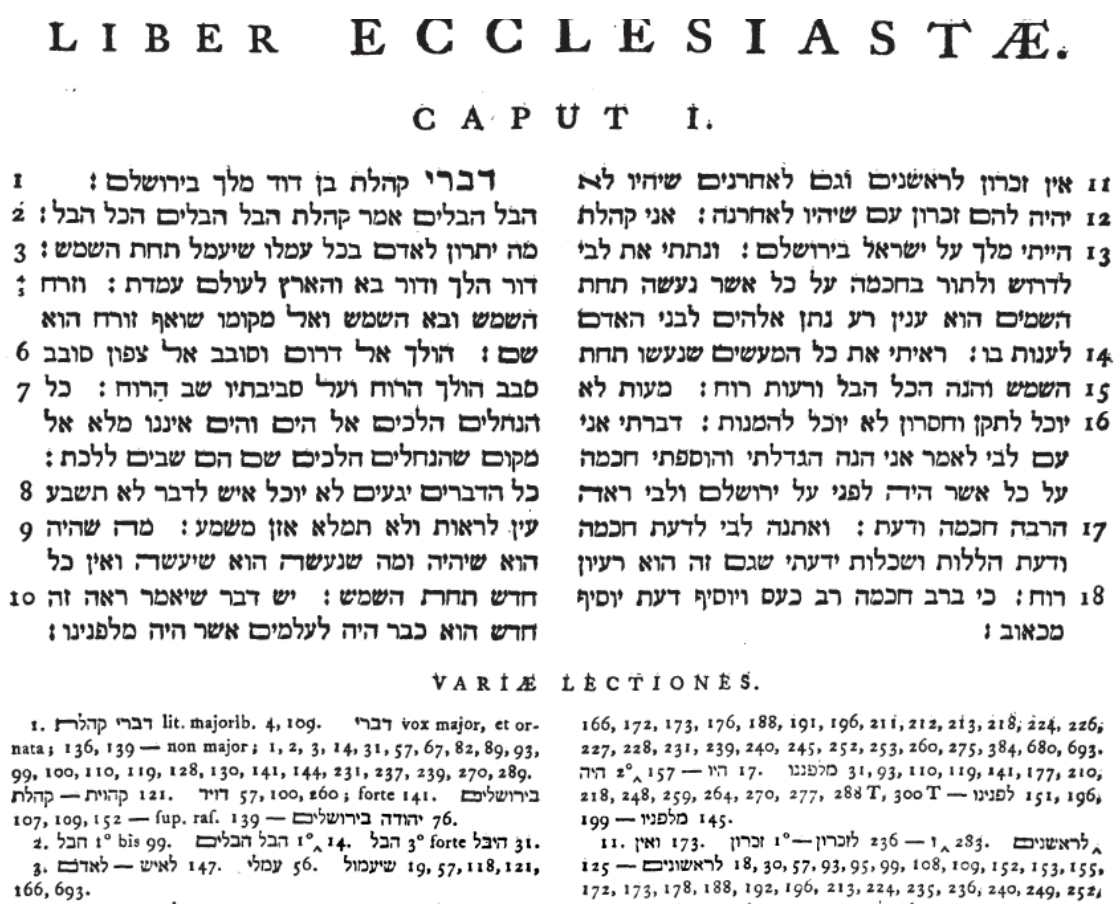
cIↃ. IↃcc. LXXXVI.

De Rossi, 1788

Medieval tradition: the XVIII c. collations

- Kennicott and De Rossi gathered thousands of variants from more than 1500 witnesses of the Hebrew Bible
- Kennicott:
  - 2 volumes, ca. 1800 pages
  - 600 witnesses
  - 1.500.000 pieces of textual information (Barthélemy)
- The collations contain important text-critical and linguistic information about the history of the biblical text and Hebrew language in the Middle Ages
- No extensive collations have been planned since then
- They are available only in a digitized format (`.pdf`)

Qohelet according to Kennicott



LIBER ECCLESIASTÆ.

CAPUT I.

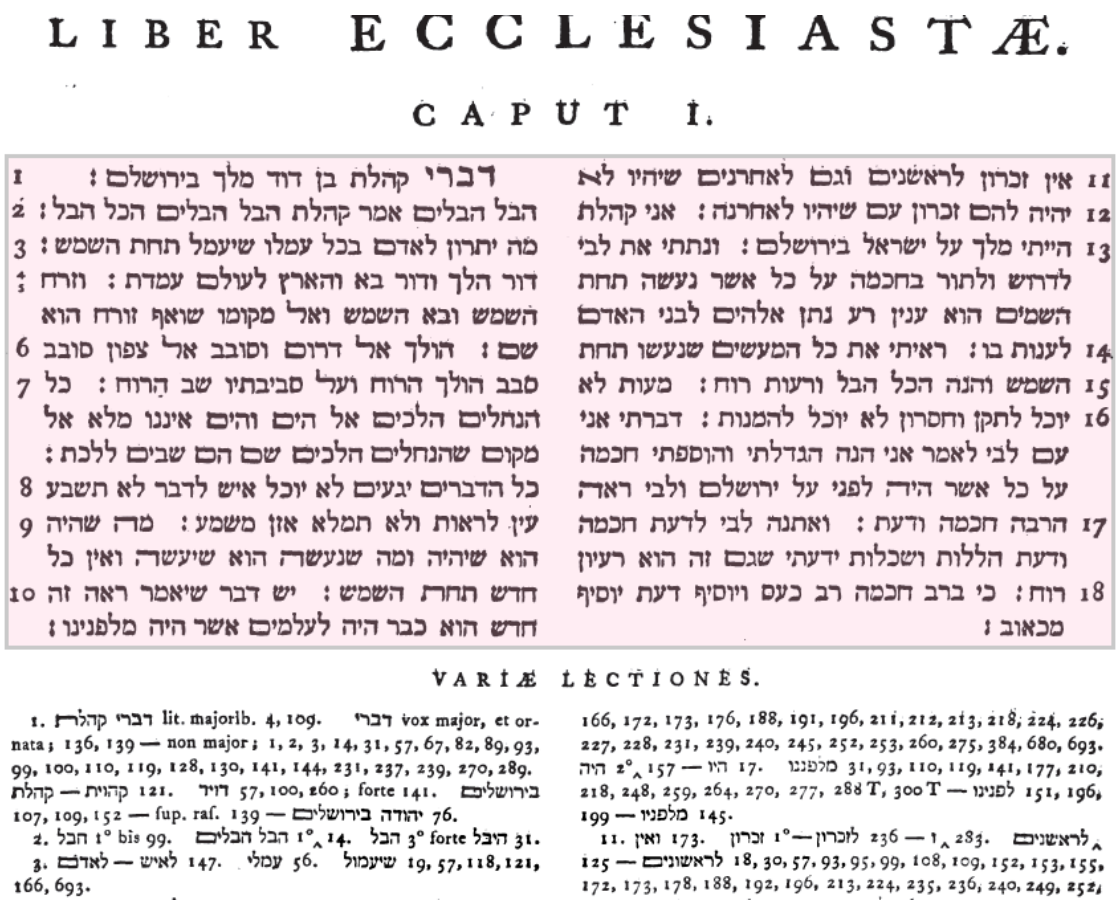
1 דברי קהלת בן דוד מלך בירושלם׃
2 הבל הבלים אמר קהלת הבל הבלים הכל הבל׃
3 מה יתרון לאדם בכל עמלו שיעמל תחת השמש׃
4 דור הלך ודור בא והארץ לעולם עמדת׃
5 וזרח השמש ובא השמש ואל מקומו שואף זורח הוא
שם׃
6 הולך אל דרום וסובב אל צפון סובב
7 סבב הולך הרוח ועל סביבתיו שב הרוח׃ כל
הנחלים הלכים אל הים והים איננו מלא אל
מקום שהנחלים הלכים שם הם שבים ללכת׃
8 כל הדברים יגעים לא יוכל איש לדבר לא תשבע
עין לראות ולא תמלא אזן משמע׃
9 מה שהיה
הוא שיהיה ומה שנעשה הוא שיעשה ואין כל
חדש תחת השמש׃
10 יש דבר שיאמר ראה זה
חדש הוא כבר היה לעלמים אשר היה מלפנינו׃
11 אין זכרון לראשנים וגם לאחרנים שיהיו לא
12 יהיה להם זכרון עם שיהיו לאחרנה׃ אני קהלת
13 הייתי מלך על ישראל בירושלם׃ ונתתי את לבי
לדרוש ולתור בחכמה על כל אשר נעשה תחת
השמים הוא ענין רע נתן אלהים לבני האדם
14 לענות בו׃ ראיתי את כל המעשים שנעשו תחת
15 השמש והנה הכל הבל ורעות רוח׃ מעות לא
16 יוכל לתקן וחסרון לא יוכל להמנות׃ דברתי אני
עם לבי לאמר אני הנה הגדלתי והוספתי חכמה
על כל אשר היה לפני על ירושלם ולבי ראה
17 הרבה חכמה ודעת׃ ואתנה לבי לדעת חכמה
ודעת הללות ושכלות ידעתי שגם זה הוא רעיון
18 רוח׃ כי ברב חכמה רב כעס ויוסיף דעת יוסיף
מכאוב׃

VARIÆ LECTIONES.

1. דברי קהלת lit. majoribus. 4, 109. דברי vox major, et ornata; 136, 139 — non major; 1, 2, 3, 14, 31, 57, 67, 82, 89, 93, 99, 100, 110, 119, 128, 130, 141, 144, 231, 237, 239, 270, 289. קהלת — קהוית 121. דויד 57, 100, 260; forte 141. בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.
2. הבל 1° bis 99. הבל הבלים 1°∧14. הבל 3° forte הבל 31.
3. לאדם — לאיש 147. עמלו 56. שיעמול 19, 57, 118, 121, 166, 693.
166, 172, 173, 176, 188, 191, 196, 211, 212, 213, 218, 224, 226, 227, 228, 231, 239, 240, 245, 252, 253, 260, 275, 384, 680, 693. היה 2°∧157 — היו 17. מלפנינו 31, 93, 110, 119, 141, 177, 210, 218, 248, 259, 264, 270, 277, 288 T, 300 T — לפנינו 151, 196, 199 — מלפניו 145.
11. ואין 173. זכרון 1° — לזכרון 236 — 1∧283. לראשנים∧ 125 — לראשונים 18, 30, 57, 93, 95, 99, 108, 109, 152, 153, 155, 172, 173, 178, 188, 192, 196, 213, 224, 235, 236, 240, 249, 252,

A Rule-Based Parsing System for Encoding Kennicott's Collation of the Hebrew Bible

Qohelet according to Kennicott



LIBER ECCLESIASTÆ.

CAPUT I.

1 דברי קהלת בן דוד מלך בירושלם׃
2 הבל הבלים אמר קהלת הבל הבלים הכל הבל׃
3 מה יתרון לאדם בכל עמלו שיעמל תחת השמש׃
4 דור הלך ודור בא והארץ לעולם עמדת׃ וזרח
השמש ובא השמש ואל מקומו שואף זורח הוא
6 שם׃ הולך אל דרום וסובב אל צפון סובב
7 סבב הולך הרוח ועל סביבתיו שב הרוח׃ כל
הנחלים הלכים אל הים והים איננו מלא אל
מקום שהנחלים הלכים שם הם שבים ללכת׃
8 כל הדברים יגעים לא יוכל איש לדבר לא תשבע
9 עין לראות ולא תמלא אזן משמע׃ מה שהיה
הוא שיהיה ומה שנעשה הוא שיעשה ואין כל
10 חדש תחת השמש׃ יש דבר שיאמר ראה זה
חדש הוא כבר היה לעלמים אשר היה מלפנינו׃
11 אין זכרון לראשנים וגם לאחרנים שיהיו לא
12 יהיה להם זכרון עם שיהיו לאחרנה׃ אני קהלת
13 הייתי מלך על ישראל בירושלם׃ ונתתי את לבי
לדרוש ולתור בחכמה על כל אשר נעשה תחת
השמים הוא ענין רע נתן אלהים לבני האדם
14 לענות בו׃ ראיתי את כל המעשים שנעשו תחת
15 השמש והנה הכל הבל ורעות רוח׃ מעות לא
16 יוכל לתקן וחסרון לא יוכל להמנות׃ דברתי אני
עם לבי לאמר אני הנה הגדלתי והוספתי חכמה
על כל אשר היה לפני על ירושלם ולבי ראה
17 הרבה חכמה ודעת׃ ואתנה לבי לדעת חכמה
ודעת הללות ושכלות ידעתי שגם זה הוא רעיון
18 רוח׃ כי ברב חכמה רב כעס ויוסיף דעת יוסיף
מכאוב׃

VARIÆ LECTIONES.

1. דברי קהלת lit. majorib. 4, 109. דברי vox major, et ornata; 136, 139 — non major; 1, 2, 3, 14, 31, 57, 67, 82, 89, 93, 99, 100, 110, 119, 128, 130, 141, 144, 231, 237, 239, 270, 289. קהלת — קהוית 121. דויד 57, 100, 260; forte 141. בירושלים 107, 109, 152 — ſup. raſ. 139 — יהודה בירושלים 76.
2. הבל 1° bis 99. הבל הבלים 1°^ 14. הבל 3° forte היבל 31.
3. לאדם — לאיש 147. עמלי 56. שיעמול 19, 57, 118, 121, 166, 693.
166, 172, 173, 176, 188, 191, 196, 211, 212, 213, 218, 224, 226, 227, 228, 231, 239, 240, 245, 252, 253, 260, 275, 384, 680, 693. היה 2°^ 157 — היו 17. מלפנינו 31, 93, 110, 119, 141, 177, 210, 218, 248, 259, 264, 270, 277, 288 T, 300 T — לפנינו 151, 196, 199 — מלפנין 145.
11. ואין 173. זכרון 1° — לזכרון 236 — 1^ 283. לראשנים^ 125 — לראשונים 18, 30, 57, 93, 95, 99, 108, 109, 152, 153, 155, 172, 173, 178, 188, 192, 196, 213, 224, 235, 236, 240, 249, 252,

- Reference text: edition of Ev. van der Hooght (1710)

Qohelet according to Kennicott

LIBER ECCLESIASTÆ.

CAPUT I.

1 דברי קהלת בן דוד מלך בירושלם׃
2 הבל הבלים אמר קהלת הבל הבלים הכל הבל׃
3 מה יתרון לאדם בכל עמלו שיעמל תחת השמש׃
4 דור הלך ודור בא והארץ לעולם עמדת׃ וזרח
השמש ובא השמש ואל מקומו שואף זורח הוא
6 שם׃ הולך אל דרום וסובב אל צפון סובב
7 סבב הולך הרוח ועל סביבתיו שב הרוח׃ כל
הנחלים הלכים אל הים והים איננו מלא אל
מקום שהנחלים הלכים שם הם שבים ללכת׃
8 כל הדברים יגעים לא יוכל איש לדבר לא תשבע
9 עין לראות ולא תמלא אזן משמע׃ מה שהיה
הוא שיהיה ומה שנעשה הוא שיעשה ואין כל
10 חדש תחת השמש׃ יש דבר שיאמר ראה זה
חדש הוא כבר היה לעלמים אשר היה מלפנינו׃
11 אין זכרון לראשנים וגם לאחרנים שיהיו לא
12 יהיה להם זכרון עם שיהיו לאחרנה׃ אני קהלת
13 הייתי מלך על ישראל בירושלם׃ ונתתי את לבי
לדרוש ולתור בחכמה על כל אשר נעשה תחת
השמים הוא ענין רע נתן אלהים לבני האדם
14 לענות בו׃ ראיתי את כל המעשים שנעשו תחת
15 השמש והנה הכל הבל ורעות רוח׃ מעות לא
16 יוכל לתקן וחסרון לא יוכל להמנות׃ דברתי אני
עם לבי לאמר אני הנה הגדלתי והוספתי חכמה
על כל אשר היה לפני על ירושלם ולבי ראה
17 הרבה חכמה ודעת׃ ואתנה לבי לדעת חכמה
ודעת הללות ושכלות ידעתי שגם זה הוא רעיון
18 רוח׃ כי ברב חכמה רב כעס ויוסיף דעת יוסיף
מכאוב׃

VARIÆ LECTIONES.

1. דברי קהלת lit. majoribus. 4, 109. דברי vox major, et ornata; 136, 139 — non major; 1, 2, 3, 14, 31, 57, 67, 82, 89, 93, 99, 100, 110, 119, 128, 130, 141, 144, 231, 237, 239, 270, 289. קהלת — קהילת 121. דויד 57, 100, 260; forte 141. בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.
2. הבל 1° bis 99. הבל הבלים 1°^ 14. הבל 3° forte הביל 31.
3. לאדם — לאיש 147. עמלי 56. שיעמול 19, 57, 118, 121, 166, 693.
166, 172, 173, 176, 188, 191, 196, 211, 212, 213, 218, 224, 226, 227, 228, 231, 239, 240, 245, 252, 253, 260, 275, 384, 680, 693. היה 2°^ 157 — היו 17. מלפנינו 31, 93, 110, 119, 141, 177, 210, 218, 248, 259, 264, 270, 277, 288 T, 300 T — לפנינו 151, 196, 199 — מלפנין 145.
11. ואין 173. זכרון 1° — לזכרון 236 — 1^ 283. לראשנים^ 125 — לראשונים 18, 30, 57, 93, 95, 99, 108, 109, 152, 153, 155, 172, 173, 178, 188, 192, 196, 213, 224, 235, 236, 240, 249, 252,

- Reference text: edition of Ev. van der Hooght (1710)
- Critical apparatus: 350 witnesses and ca. 2600 variants

Digitalizing Kennicott's collation

Digitalization can provide useful information for:

- **Textual history** and **philology**
  - Computing genealogical relationships (stemmatology)

    [Hempel, Goshen-Gottstein, Gese, Sacchi, Borbone]
  - Studying the phenomenology of the copying process (scribal habits, genesis of common copying errors...)
- **Codicology** and **paleography**
  - Classifying mss according to ethno-geographic criteria (Ashkenazic/Sephardic/Italian mss...)

    [Penkower 1988, 2002]
- **Linguistics**
  - *usus scribendi* (*scriptio plena*/*defectiva*, orthography...)

    [Cohen 1986]

Digitalizing Kennicott's collation

In order to permit the computer to extract information from the critical apparatus, data need to be not only machine readable, but also fully machine actionable

Machine readability is achievied through **OCR** technology (digitization)

Machine actionability is ensured by textual encoding through markup languages (**XML**)

Towards an automated encoding

- Manually encoding critical apparatus is expensive, time-consuming and error-prone
- It is possible to encode automatically through Natural Language Processing (**NLP**) tools
- There are two main approaches:
  1. **Rule-based** systems:
     - Rules for describing the language are defined by the user
     - Robust, but the language needs to be as structured as possible
     - Relatively simple
  2. **Machine learning** systems:
     - Rules for describing the language are derived from the data (machine learning algorithms)
     - More efficient for non- or semi-structured languages
     - Complex

The language of Kennicott's apparatus

- Well structured
- Textual phenomena are expressed in a non-redundant and unambigous way by means of:

1. Conventional vocabulary: **set of finite symbols** (numbers, strings, abbreviations)
2. Rigorous syntax: the **position** of apparatus components conveys information about their function

Parsing the critical apparatus

An example from Qoh. 1:1

1. דברי קהלת בן דוד מלך בירושלם

Words of Qohelet, son of David king in Jerusalem

Kennicott's apparatus

1. **דברי קהלת** lit. majorib. 4, 109.

קהלת — קהוית 121.

דויד 57, 100, 260 ; forte 141.

בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.

Apparatus components

1. דברי קהלת בן דוד מלך בירושלם

1. **דברי קהלת** lit. majorib. 4, 109.

קהלת — **קהוית** 121.

דויד 57, 100, 260 ; forte 141.

בירושלים 107, 109, 152 — sup. ras. 139 — **יהודה בירושלים** 76.

Apparatus components

1. דברי קהלת בן דוד מלך בירושלם

1. דברי קהלת lit. majorib. 4, 109.

קהלת — קהוית 121.

דויד 57, 100, 260 ; forte 141.

בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.

variation place (number of chapter and verse)

Apparatus components

1. דברי קהלת בן דוד מלך בירושלם

1. דברי קהלת lit. majorib. 4, 109.

קהלת — קהות 121.

דויד 57, 100, 260 ; forte 141.

בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.

readings

Apparatus components

1. דברי קהלת בן דוד מלך בירושלם

1. דברי קהלת lit. majorib. 4, 109.

קהלת — קהות 121.

דויד 57, 100, 260 ; forte 141.

בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.

witnesses sigla

Apparatus components

1. דברי קהלת בן דוד מלך בירושלם

1. דברי קהלת lit. majorib. 4, 109.

קהלת — קהוית 121.

דויד 57, 100, 260 ; forte 141.

בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.

variant description (*primo*, *nunc*, *forte* etc.)

Apparatus components

1. דברי קהלת בן דוד מלך בירושלם

1. דברי קהלת lit. majorib. 4 , 109 .

קהלת — קהוית121 .

דויד 57 , 100 , 260 ; forte 141 .

בירושלים 107 , 109 , 152 — sup. ras. 139 — יהודה בירושלים76 .

separators

Apparatus components

1. דברי קהלת בן דוד מלך בירושלם

1. דברי קהלת lit. majorib. 4 , 109 .

קהלת — קהות121 .

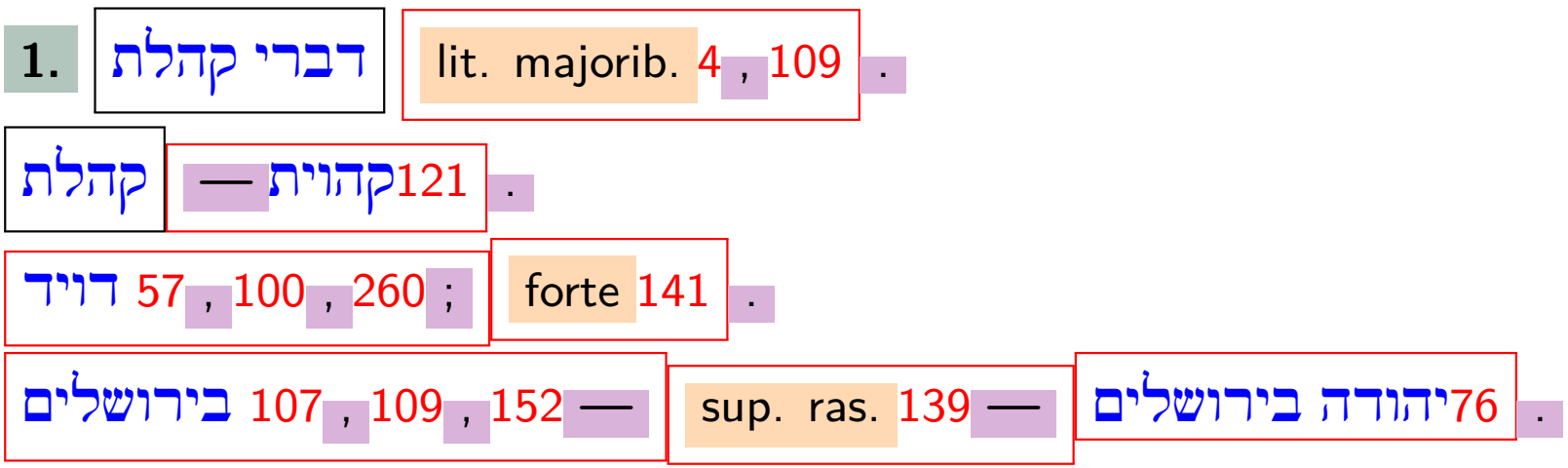
דויד 57 , 100 , 260 ; forte 141 .

בירושלים 107 , 109 , 152 — sup. ras. 139 — יהודה בירושלים76 .

lemma (reading from the reference text)

Apparatus components

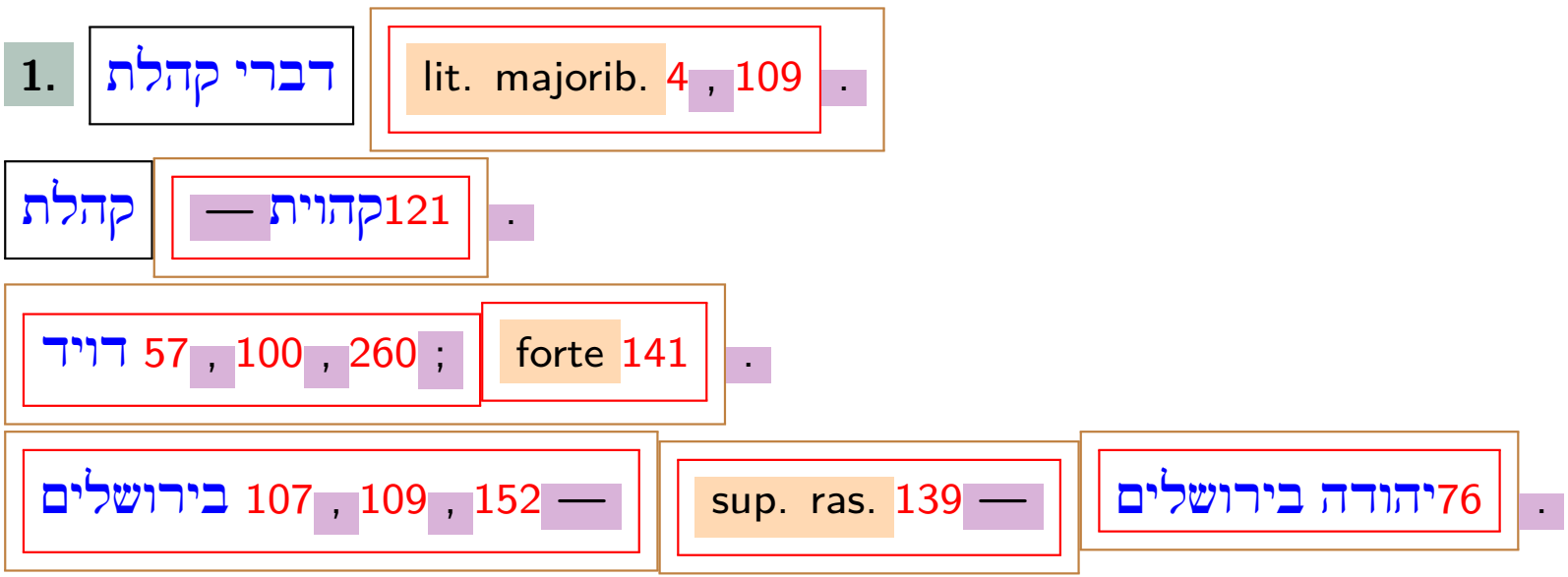
1. דברי קהלת בן דוד מלך בירושלם



variant readings diverging from the lemma

Apparatus components

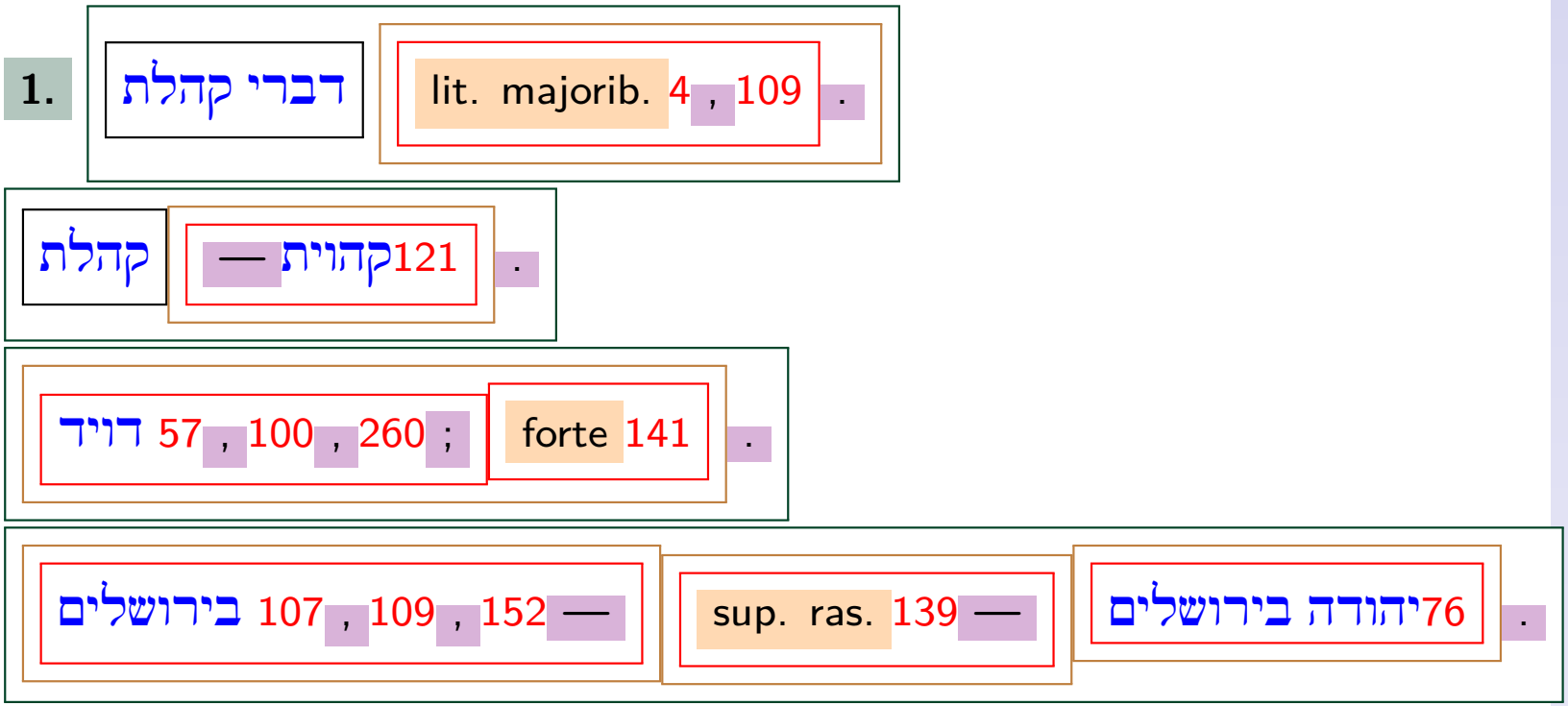
1. דברי קהלת בן דוד מלך בירושלם



reading groups

Apparatus components

1. דברי קהלת בן דוד מלך בירושלם



apparatus entries

A rule-based parsing system

- It is possible to exploit the nature of structured language characterizing Kennicott's apparatus in order to perform an automated encoding
- The machine is instructed on how to recognize the apparatus components through a parser
- A parser is a software that analyses sequences of strings (parsing) according to given rules (rule-based)

A rule-based parsing system: pipeline

1. The apparatus is processed through OCR technology to make it machine readable
2. A context-free grammar (**CFG**) is built to describe the language of the apparatus
3. A Visitor is implemented to produce XML code
4. **XSL-T** stylesheets convert XML to TEI

A rule-based parsing system: pipeline

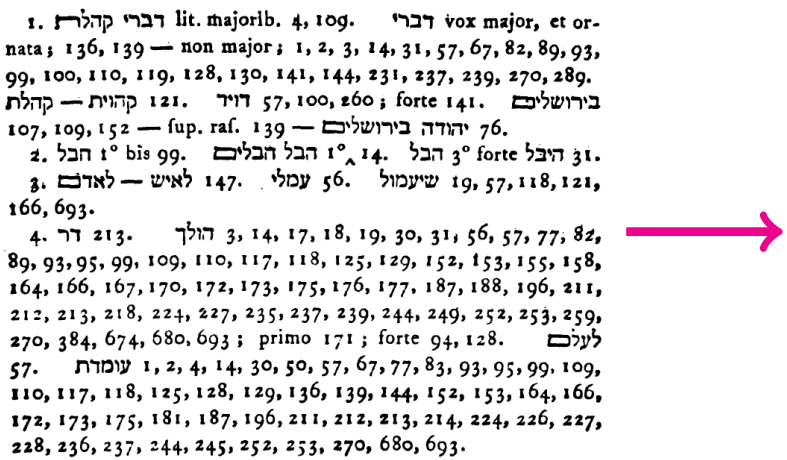
1. The apparatus is processed through OCR technology to make it machine readable
2. A context-free grammar (**CFG**) is built to describe the language of the apparatus
3. A Visitor is implemented to produce XML code
4. **XSL-T** stylesheets convert XML to TEI

ANTLR 4 (**AN***other* **T***ool for* **L***anguage* **R***ecognition*)

https://www.antlr.org/

1. Optical character recognition

- Data are converted to machine readable format (`.txt`) through Tesseract (https://github.com/tesseract-ocr/tesseract)
- The output is corrected by hand



1. דברי קהלת lit. majorib. 4, 109. דברי vox major, et ornata ; 136, 139 — non major 141, 144, 231, 237, 239, 270, 289. קהלת — קהוית 121. דויד 57, 100, 260 ; forte 2. הבל 1° bis 99. הבל הבלים 1° . 14. הבל 3° forte היבל 31. 3. לאדם — לאיש 147. עמלי 56. שיעמול 19, 57, 118, 121, 166, 693. 4. דר 213. הולך 3, 14, 17, 18, 19, 30, 31, 56, 57, 77, 82, 89, 93, 95, 99, 172, 173, 175, 176, 177, 187, 188, 196, 211, 212, 213, 218, 224, 227, 235, 237, forte 94, 128. לעלם 57. עומדת 1, 2, 4, 14, 30, 50, 57, 67, 77, 83, 93, 95, 99, 1 172, 173, 175, 181, 187, 196, 211, 212, 213, 214, 224, 226, 227, 228, 236, 237, 5. ובא השמש . 170, 178. ואל — ו . 18, 151, 152, 198, 201, 218, 224. מקמו 2, 223, 231, 384. שאף 2, 3, 17, 77, 89, 111, 118, 121, 151, 152, 155, 157, 158, 1 111, 121, 129, 158, 172, 187, 213, 218, 223, 224, 236, 237, 239, 244 — וזורח 240. 6. הולך 1° — ו . 18, 77, 107, 108, 151, 155, 176. וסבב 18, 30, 101, 107, 117, 141, 117, 141, 166, 187, 212, 213, 223, 231, 239, 240, 244, 248, 253, 260. סובב 1, 166, 167, 172, 196, 211, 218, 224, 235, 237, 252, 384 ; primo 4. הולך 2° הרוח 14, 167. ועל — primo ואל 107 — ו . 152, 198. סביבותיו 1 — סבבתיו 56, 57, 93, 108, 693 ; primo 128. שם — שב 2 ; primo 50 ; videtur 136. ועל--הרוח . 170. 7. הלכים 1° — הולכים 2, 4, 17, 56, 77, 95, 99, 111, 121, 125, 129, 150, 152, 155, מלא--הים . 157. אל 2° . 107. הלכים 2° — הולכים 2, 17, 56, 77, 95, 99, 111, 119 151. 8. לא 1° — ולא 201, 674. יוכל — forte אוכל 178. איש . 19, 129 — י . 259. לא 2° — ו . 147. אוזן 152, 680 — . 384. משמע . 76 — משמוע 2, 17, 57, 150, 158, 164, 172, 9. שנעש' 121. אין 18, 30, 82, 93, 95, 117, 118, 125, 129, 147, 151, 152, 158,

2. The Context-free grammar

- The Context-free grammar (CFG) is a formal grammar consisting of a set of top-down, rewriting rules
- The rules of a CFG describe a formal language
- A lexer and a parser are implemented in ANTLR4 on the base of the CFG
- The lexer tokenizes the formal language (tokenization rules)
- The parser checks its syntax (parser rules) creating an Abstract Syntactic Tree (**AST**)

2. The Context-free grammar

The readings

דברי קהלת lit. majorib. 4, 109.

2. The Context-free grammar

The readings

דברי קהלת lit. majorib. 4, 109.

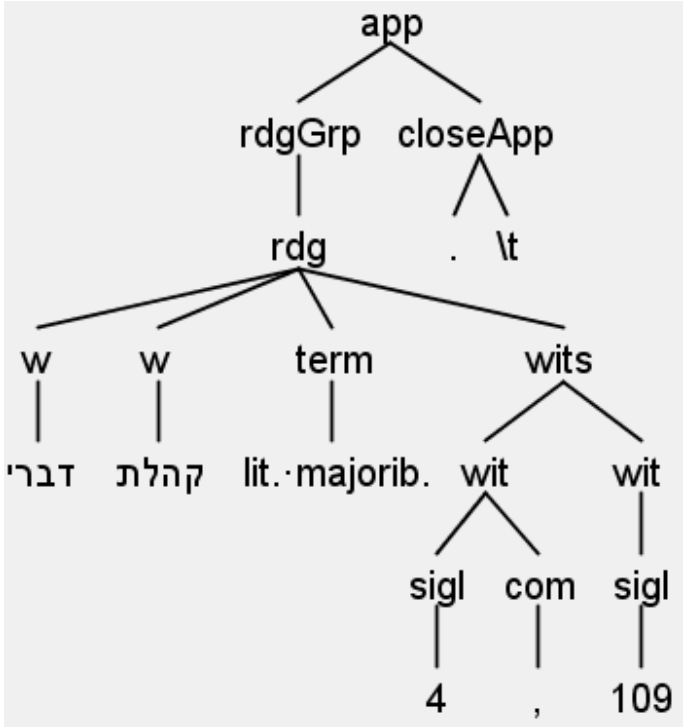
```
7   grammar kennicottCFG;
8   all: listApp+;
9   listApp: loc app+ ;
10  app: lem?  rdgGrp+ closeApp;
11  lem: w+ lemSep;
12  rdgGrp: (rdg+) rdgGrpSep?;
13  rdg: (w+)? (term)? wits rdgSep?;
14  w: HEBW;
15  loc: verse closeLoc;
16  term: MAN_DESC;
17  wits : wit+;
18  sigl: NUM;
19  wit: sigl com?;
20  lemSep: VAR_SEP;
21  rdgGrpSep: VAR_SEP ;
22  com: COMMA;
23  rdgSep: COMMA | SEMICOLON;
24  numSign: NUMEROSIGN ;
25  verse: num+ ;
26  num: NUM;
27  closeMainApp: NEWLINE;
28  closeApp: END TAB | END NEWLINE ;
29  closeLoc: END;
30
31  MAN_DESC: 'forte' | 'sup. ras.' |
32  'lit. majorib.' ;
33  ALPHA_SEQ : [a-zA-Z]+;
34  NUM : [0-9]+('.'[0-9]+)?;
35  HEBW : [\u0590-\u05ff*]+;
36  NEWLINE: ('\n');
```

2. The Context-free grammar

The readings

דברי קהלת lit. majorib. 4, 109.

```
7   grammar kennicottCFG;
8   all: listApp+;
9   listApp: loc app+ ;
10  app: lem?  rdgGrp+ closeApp;
11  lem: w+ lemSep;
12  rdgGrp: (rdg+) rdgGrpSep?;
13  rdg: (w+)? (term)? wits rdgSep?;
14  w: HEBW;
15  loc: verse closeLoc;
16  term: MAN_DESC;
17  wits : wit+;
18  sigl: NUM;
19  wit: sigl com?;
20  lemSep: VAR_SEP;
21  rdgGrpSep: VAR_SEP ;
22  com: COMMA;
23  rdgSep: COMMA | SEMICOLON;
24  numSign: NUMEROSIGN ;
25  verse: num+ ;
26  num: NUM;
27  closeMainApp: NEWLINE;
28  closeApp: END TAB | END NEWLINE ;
29  closeLoc: END;
30
31  MAN_DESC: 'forte' | 'sup. ras.' |
32  'lit. majorib.' ;
33  ALPHA_SEQ : [a-zA-Z]+;
34  NUM : [0-9]+('.'[0-9]+)?;
35  HEBW : [\u0590-\u05ff*]+;
36  NEWLINE: ('\n');
```



2. The Context-free grammar

The reading groups

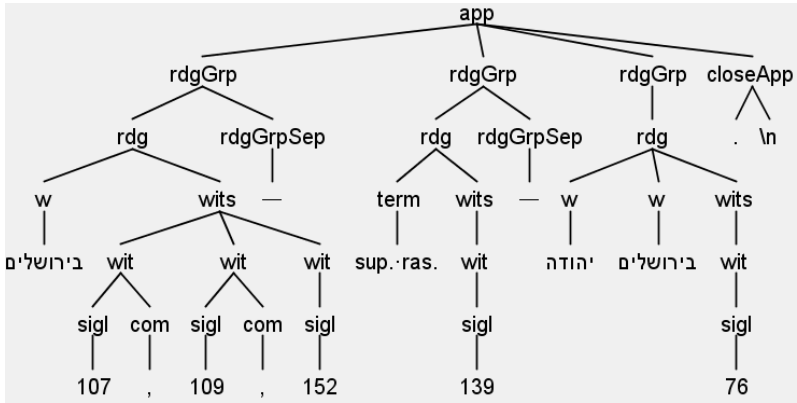
בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.

2. The Context-free grammar

The reading groups

בירושלים 107, 109, 152 — sup. ras. 139 — יהודה בירושלים 76.

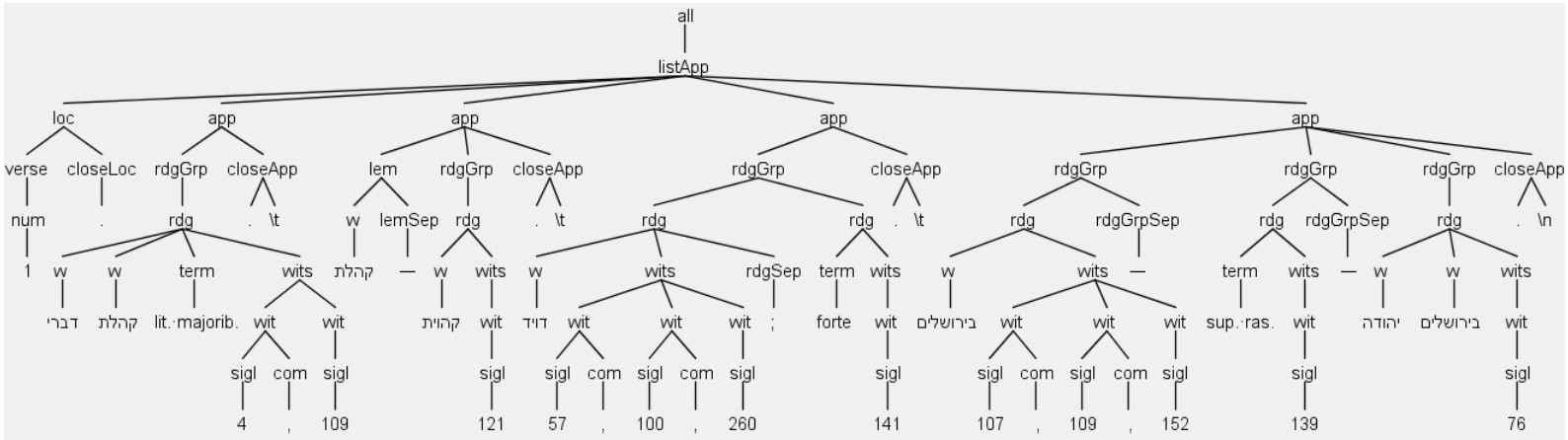
```
7   grammar kennicottCFG;
8   all: listApp+;
9   listApp: loc app+ ;
10  app: lem?  rdgGrp+ closeApp;
11  lem: w+ lemSep;
12  rdgGrp: (rdg+) rdgGrpSep?;
13  rdg: (w+)? (term)? wits rdgSep?;
14  w: HEBW;
15  loc: verse closeLoc;
16  term: MAN_DESC;
17  wits : wit+;
18  sigl: NUM;
19  wit: sigl com?;
20  lemSep: VAR_SEP;
21  rdgGrpSep: VAR_SEP ;
22  com: COMMA;
23  rdgSep: COMMA | SEMICOLON;
24  numSign: NUMEROSIGN ;
25  verse: num+ ;
26  num: NUM;
27  closeMainApp: NEWLINE;
28  closeApp: END TAB | END NEWLINE ;
29  closeLoc: END;
30
31  MAN_DESC: 'forte' | 'sup. ras.' |
32  'lit. majorib.' ;
33  ALPHA_SEQ : [a-zA-Z]+;
34  NUM : [0-9]+('.'[0-9]+)?;
35  HEBW : [\u0590-\u05ff*]+;
36  NEWLINE: ('\n');
```



2. The Context-free grammar

The apparatus entry

1. דברי קהלת lit. majorib. 4, 109. קהלת – קהות 121. דויד 57, 100, 260 ; forte 141. בירושלים 107, 109, 152 – sup. ras. 139 – יהודה בירושלים 76.



3. The Visitor

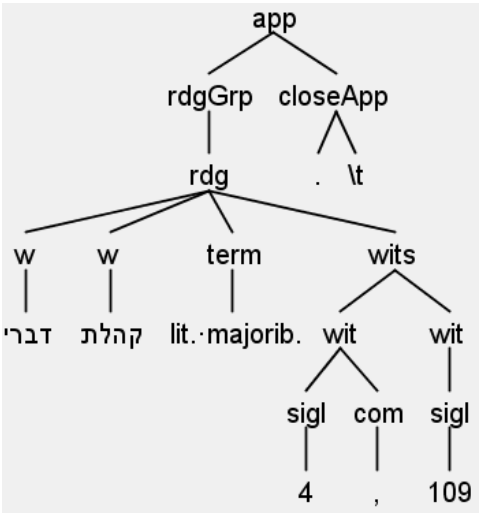
- The Visitor is a tree-walking mechanism
- In this implementation, the Visitor traverses the tree and slavishly translates the CFG rules into XML nodes
- It is a general-purpose exporter: once implemented, it can be applied to every CFG without further customization

3. The Visitor

- The Visitor is a tree-walking mechanism
- In this implementation, the Visitor traverses the tree and slavishly translates the CFG rules into XML nodes
- It is a general-purpose exporter: once implemented, it can be applied to every CFG without further customization

https://cophilab.ilc.cnr.it/parseForge/
(Courtesy of F. Boschetti)

3. The Visitor: from the AST to XML




```
1  <listapp>
2    <loc>
3      <verse>
4        <num>1</num>
5      </verse>
6      <closeLoc>.</closeLoc>
7    </loc>
8    <app>
9      <rdgGrp>
10       <rdg>
11         <w>דברי</w>
12         <w>קהלת</w>
13         <term>lit. majorib.</term>
14         <wit>
15           <sigl>4</sigl>
16           <com>,</com>
17         </wit>
18         <wit>
19           <sigl>109</sigl>
20         </wit>
21       </rdg>
22     </rdgGrp>
23   </app>
24 </listapp>
25 ...
```

4. XSL-T: from XML to TEI

Location referenced method

```
1  <listapp>
2    <loc>
3      <verse>
4        <num>1</num>
5      </verse>
6      <closeLoc>.</closeLoc>
7    </loc>
8    <app>
9      <rdgGrp>
10       <rdg>
11         <w>דברי</w>
12         <w>קהלת</w>
13         <term>lit. majorib.</term>
14         <wit>
15           <sigl>4</sigl>
16           <com>,</com>
17         </wit>
18         <wit>
19           <sigl>109</sigl>
20         </wit>
21       </rdg>
22     </rdgGrp>
23   </app>
24 </listapp>
25 ...
```

```
1  <listapp>
2    <app loc="1 1">
3      <lem>
4        <w>דברי</w>
5        <w>קהלת</w>
6      </lem>
7      <rdgGrp>
8        <rdg wit="#K4 #K109">
9          <term>lit. majorib.</term>
10       </rdg>
11     </rdgGrp>
12   </app>
13   <app loc="1 1">
14     <lem>
15       <w>קהלת</w>
16     </lem>
17     <rdgGrp>
18       <rdg wit="#K121">
19         <w>קהוית</w>
20       </rdg>
21     </rdgGrp>
22   </app>
23   ...
```

A rule-based parser

Results

1. 100% of accuracy for Qohelet

A rule-based parser

Results

1. 100% of accuracy for Qohelet
2. 2617 variants correctly parsed and encoded

A rule-based parser

Results

1. 100% of accuracy for Qohelet
2. 2617 variants correctly parsed and encoded
3. Rut (1025 variants) and Song of Songs (1238): few syntactic errors (99% accuracy)

A rule-based parser

Results

1. 100% of accuracy for Qohelet
2. 2617 variants correctly parsed and encoded
3. Rut (1025 variants) and Song of Songs (1238): few syntactic errors (99% accuracy)
4. The CFG is extended in order to include unseen textual phenomena (100% accuracy)

A rule-based parser

Evaluation

- The system is robust
- Other biblical books of Kennicott's collection can be encoded

A rule-based parser

Evaluation

- The system is robust
- Other biblical books of Kennicott's collection can be encoded
- A parser approach is faster than manual encoding
- It is possible to export data in different formats (XML, HTML, relational databases, LaTeXetc.)
- It is less error prone: the CFG provides a spell-checking system, useful for detecting
  1. errors generated after OCR (transcriptional errors)
  2. inconsistencies in the printed source
- Tighter control on semantic errors when exporting to XML-TEI

Treatment of residuals

- Kennicott's apparatus minimizes natural language
- Encoding by means of concise and standard annotations

Treatment of residuals

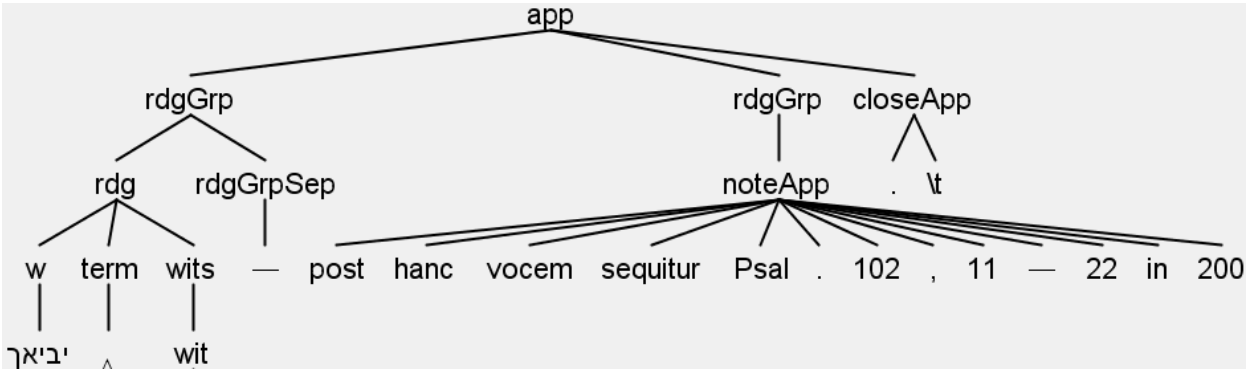
- Kennicott's apparatus minimizes natural language
- Encoding by means of concise and standard annotations
- Exceptions (residuals):
  - 7:29 "Incipit cap. 8 a voce חכמת, medio commatis..."
  - 11:9 "יביאך — post hanc vocem sequitur Psal. 102..."

Treatment of residuals

- Kennicott's apparatus minimizes natural language
- Encoding by means of concise and standard annotations
- Exceptions (residuals):
  - 7:29 "Incipit cap. 8 a voce חכמת, medio commatis..."
  - 11:9 "יביאך — post hanc vocem sequitur Psal. 102..."

Residuals are treated as unprocessed/unstructured data:

- Correctly identified (no syntactic errors)



- Stored in note-like XML elements

Stemmatic analysis

Stemmatic analysis

- Encoding can be used for quantitative analysis (e. g. computer-assisted stemmatology)

Stemmatic analysis

- Encoding can be used for quantitative analysis (e. g. computer-assisted stemmatology)
- Few attempts to study the medieval tradition of the Hebrew Bible according to (neo-)Lachmannian criteria
- Mss are treated singularly, not as groups or families
- A stemmatic classification of medieval mss is still missing

From XML-TEI to data matrix

In order to permit the computer to compute genealogical relationships, it is necessary to transform philological data into a numerical format (data matrix)

```
1  <listapp>
2    <app loc="1 1">
3      <lem>
4        <w>דברי</w>
5        <w>קהלת</w>
6      </lem>
7      <rdgGrp>
8        <rdg wit="#K4 #K109">
9          <term>lit. majorib.</term>
10       </rdg>
11     </rdgGrp>
12   </app>
13   <app loc="1 1">
14     <lem>
15       <w>קהלת</w>
16     </lem>
17     <rdgGrp>
18       <rdg wit="#K121">
19         <w>קהוית</w>
20       </rdg>
21     </rdgGrp>
22   </app>
23   ...
```

From XML-TEI to data matrix

In order to permit the computer to compute genealogical relationships, it is necessary to transform philological data into a numerical format (data matrix)

```
1  <listapp>
2    <app loc="1 1">
3      <lem>
4        <w>דברי</w>
5        <w>קהלת</w>
6      </lem>
7      <rdgGrp>
8        <rdg wit="#K4 #K109">
9          <term>lit. majorib.</term>
10       </rdg>
11     </rdgGrp>
12   </app>
13   <app loc="1 1">
14     <lem>
15       <w>קהלת</w>
16     </lem>
17     <rdgGrp>
18       <rdg wit="#K121">
19         <w>קהוית</w>
20       </rdg>
21     </rdgGrp>
22   </app>
23   ...
```

→

```
                    1         2         3         4         5
Taxon      12345678901234567890123456789012345678901234567890
-----------------------------------------------------------------
EK         00000000000000000000000000000000000000000000000000
0OL        00000000000000000000000000000000000000000000000000
3IAK1      00000000000000010001000000000000000010000000001000
3SK2       00001000000100000010000000000000000000000000000000
4SK3       00000000000000000000000010000000000000000000000000
2AK4       00000101000000000000010000000000010000000000000000
4SK14      00100000000000000010000000000000000000000000000001
3AK17      00000100000100000001001000000000001010000000000000
3AK18      01000010000100001001101000000001010100001000000000
4SK19      00000000001000000000000000000000010000000000000000
2AK30      00000010001000000100001000000000000000000000000010
0OK31      00000000000000000010000000000000000000000000000000
3AK50      00001000000000000000000000000000000000000000000000
3AK56      00000000000000000010000001000000000110100000000000
4SK67      00000000000000000000000000000000000000000000000000
2AK76      00000001000000000000001001000000001001100000000000
2AK77      00000100010100000010011110000010001100001010000000
2AK80      00000000000000000000000000000000000000000000000000
3SK82      00000010000000000001001000000000001000000000000000
3SK83      00000000000000000000000000000000000000000000000000
3SK89      00000000000000000000001000000000000000000000000000
3AK93      00000010000000000001001000000000001000000000000000
2IK94      00000000000000000000000000000000000000000000000000
2AK95      00000110000000000010100000000001010010000000000000
3SK99      00000000000000000001001000000000000000000000000000
3SK100     00000000000100000000000000000000000000000000000000
```

Data matrix: regularization

Elimination of

- Partially collated witnesses
- Accidentals (e. g. orthographic variants)
- *Marginalia*, dubious and second hand variants

Total: 119 witnesses and 371 variants

Phylogenetic analysis

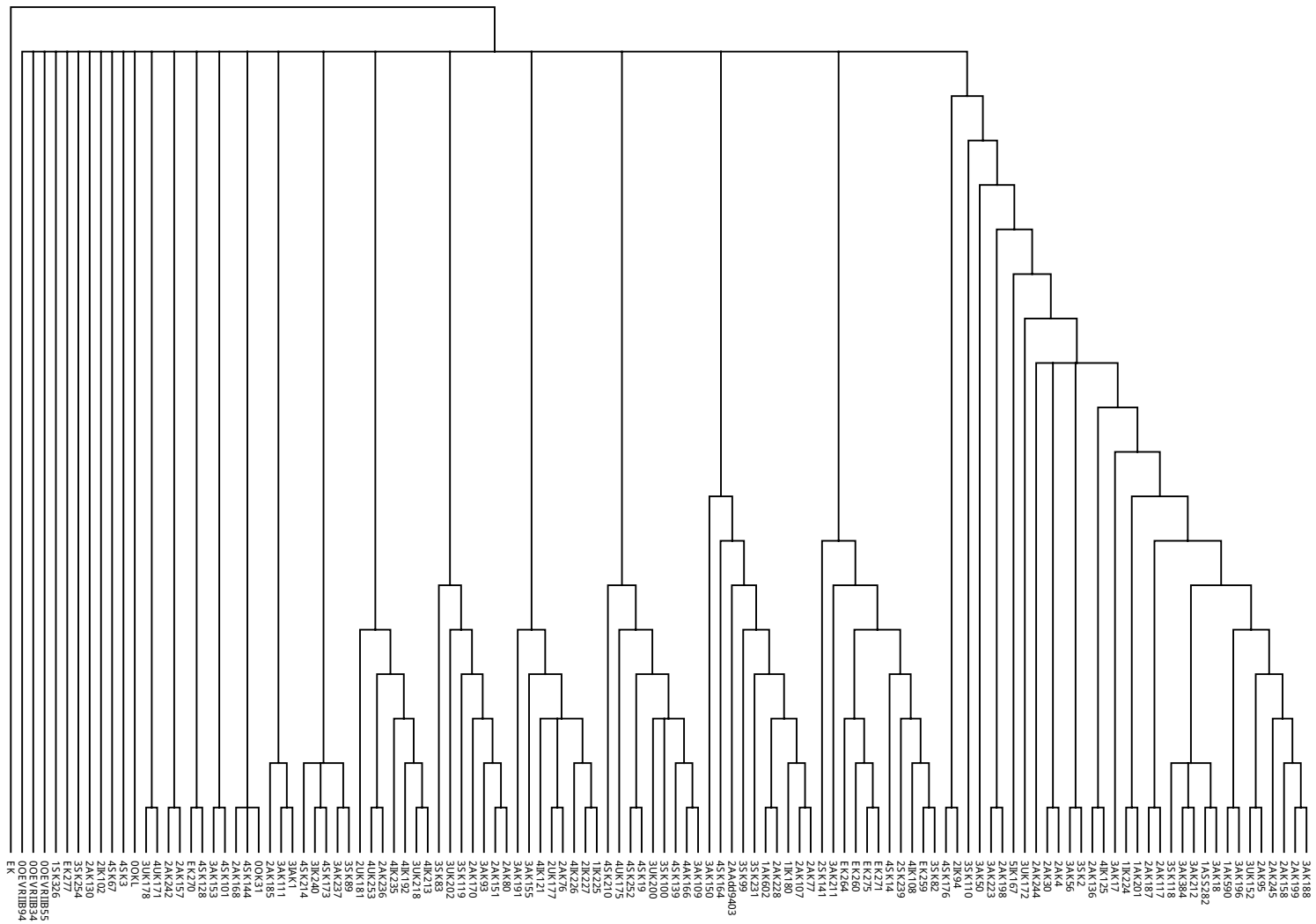
- The data matrix is analysed by phylogenetic algorithms, which produce tree-like graphs (phylograms) representing paths of textual evolution
- Many phylogenetic algorithms: Maximum Parsimony
- Implemented in **PAUP** software

P*arsimony* **A***nalysis* **U***sing* **P***AUP*

http://paup.phylosolutions.com/

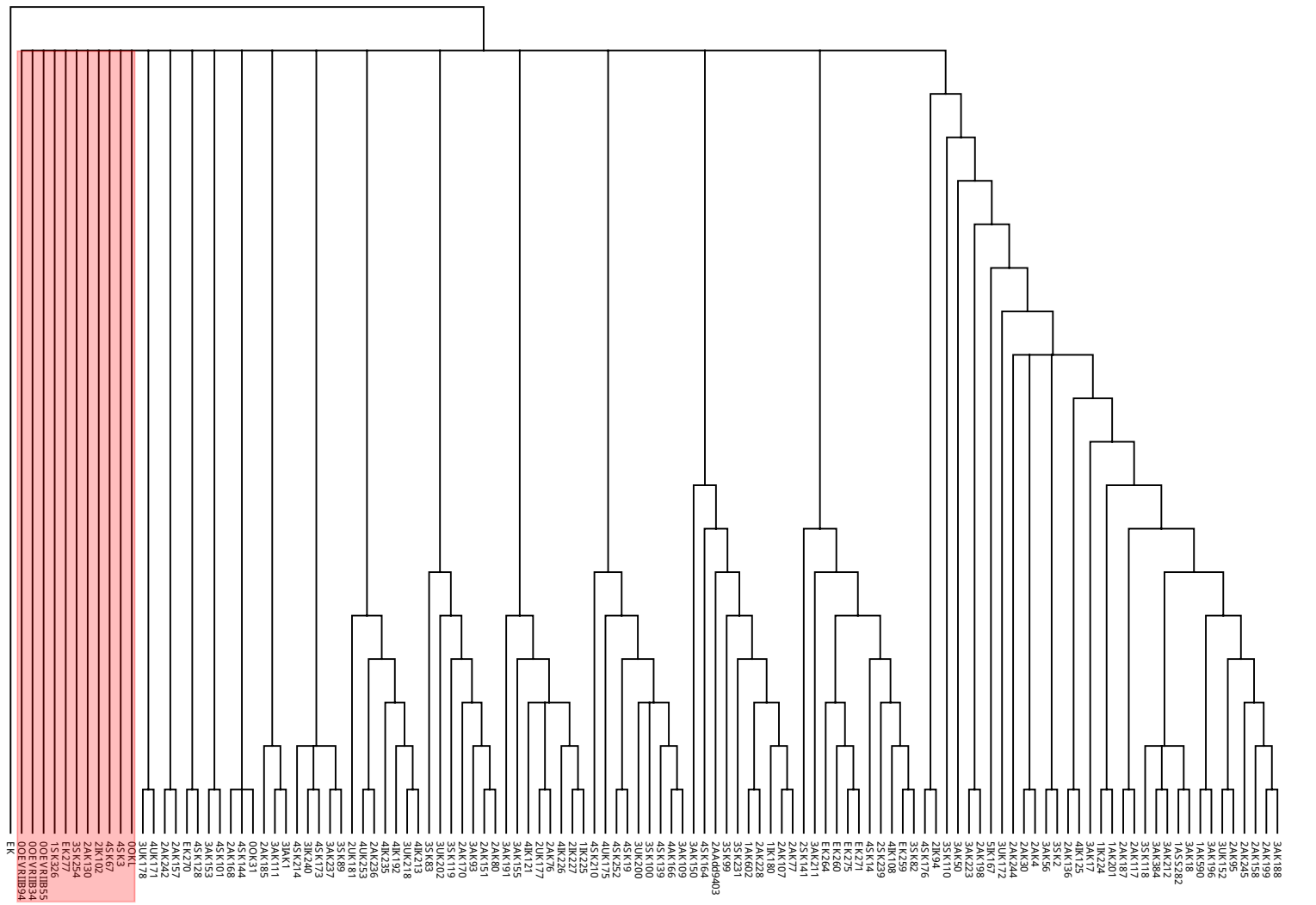
Strict consensus tree

Consensus of more than 25.000 equally parsimonious trees



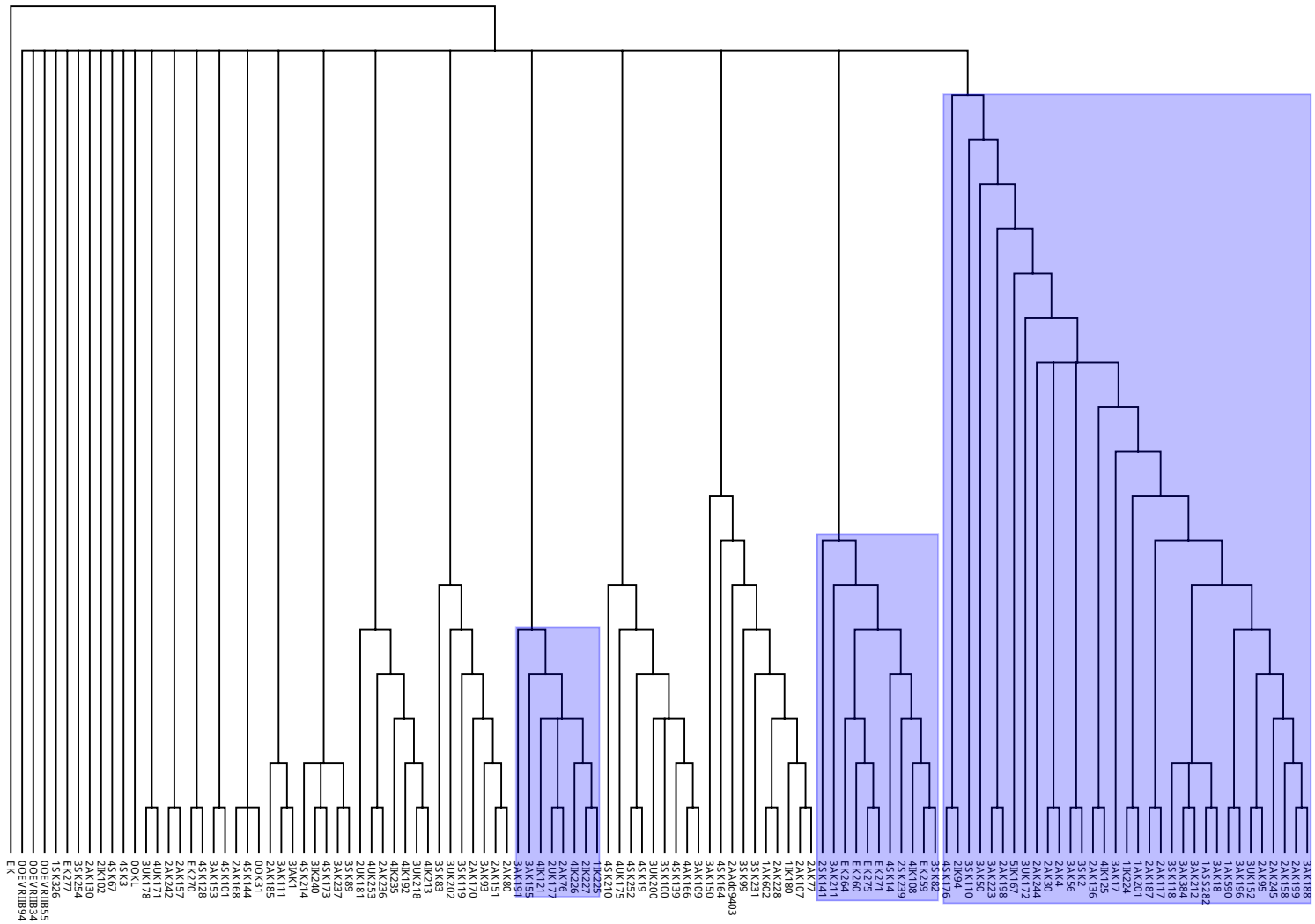
Strict consensus tree

Consensus of more than 25.000 equally parsimonious trees



Strict consensus tree

Consensus of more than 25.000 equally parsimonious trees



Phylogenetic methods and biblical mss

- Groups are identified according to ancestral variants
- Some groups have characteristic variants
- Corroborated by external criteria

Phylogenetic methods and biblical mss

- Groups are identified according to ancestral variants
- Some groups have characteristic variants
- Corroborated by external criteria

Limits

- Important mss are missing
- Variants concern consonantal text only
- Most are polygenetic
- Contamination

Phylogenetic methods and biblical mss

- Groups are identified according to ancestral variants
- Some groups have characteristic variants
- Corroborated by external criteria

Limits

- Important mss are missing
- Variants concern consonantal text only
- Most are polygenetic
- Contamination

Future works

- More data:
  - more mss
  - punctuation, Massora, para-textual variants
- Variants can be weighted and ordered

Bibliographical references

Barthélemy, D. (1992). Les manuscrits médiévaux et le texte tibérien classique. In *Critique textuelle de l'Ancien Testament, 3. Ézéchiel, Daniel et les 12 Prophètes*, Volume 3 of *Orbis Biblicus et Orientalis*, pp. xix–xcvi. Fribourg/Göttingen: Éditions Universitaires/Vandenhoeck & Ruprecht.

Borbone, P. G. (1990). *Il libro del profeta Osea. Edizione critica del testo ebraico*. Torino: Zamorani.

Cohen, M. (1973). מגבשי כתיב במצחפי מסורה עתיקים ומשמעם לתולדות נוסח המקרא המקובל. Ph. D. thesis, Hebrew University, Jerusalem. [Unpublished].

Cohen, M. (1979). האידיאה בדבר קדושת הנוסח לאותיותיו וביקורת הטכסט. In S. Uriel (Ed.), המקרא ואנחנו, Volume 1, pp. 42–69. Tel Aviv: דביר / Dvir.

Cohen, M. (1980). קווי יסוד לדמותו העיצורית של הטקסט בכתבי יד מקראיים מימי הביניים. עיוני מקרא ופרשנות / *Studies in Bible and Exegesis 1*, 123–182.

Cohen, M. (1981). לדמותם הקונסונאנטית של דפוסי המקרא הראשונים: המהדורה הראשונה של התנ"ך השלם - דפוס סונצ'ינו משנת 1488 / The Consonantal Character of the First Rabbinic Printings: the *Editio Princeps* of the Entire Bible Soncino 1488. In ספר השנה של אוניברסיטת בר־אילן, Volume XVIII-XIX, pp. 47–67. Ramat Gan: University Press.

Cohen, M. (1986). מהו 'נוסח המסורה', ומה היקף אחיזתו בתולדות המסירה של ימה"ב / The 'Masoretic Text' and the Extent of Its Influence on the Transmission of the Biblical Text in the Middle Ages. In S. Uriel (Ed.), עיוני מקרא ופרשנות / *Studies in Bible and Exegesis*, Volume 2, pp. 229–256. Ramat Gan: Bar Ilan University Press.

Gese, H. (1957). Die hebräischen Bibelhandschriften zum Dodekapropheton nach der Variantensammlung des *Kennicott*. *Zeitschrift für die Alttestamentliche Wissenschaft 69*(1–4), 55.

Goshen-Gottstein, M. H. (1954). Die Jesaiah-Rolle und das Problem der hebräischen Bibelhandschriften. *Biblica 35*(4), 429–442.

Hempel, J. (1930). Chronik. *Zeitschrift für die Alttestamentliche Wissenschaft 48*, 187–206.

Hempel, J. (1934). Innermasoretische Bestätigungen des Samaritanus. *Zeitschrift für die Alttestamentliche Wissenschaft 52*(1), 254–274.

Penkower, J. S. (1982). יעקב בן חיים וצמיחת מהדורת המקראות הגדולות / *Jacob Ben-Ḥayyim and the Rise of the Biblia Rabbinica*. Ph. D. thesis, Hebrew University, Jerusalem. [Unpublished].

Penkower, J. S. (1988). כתב-יד ירושלמי של התורה מן המאה העשירית שהגיהו מישאל בן עוזיאל (כתב-יד ק3) / A Tenth-century Pentateuchal MS from Jerusalem (MS C3), Corrected by Mishael ben Uzziel. תרביץ / *Tarbiz 58*(1), 49–74.

Penkower, J. S. (2002). A Sheet of Parchment from a 10th or 11th Century Torah Scroll: Determining Its Type Among Four Traditions (Oriental, Sefardi, Ashkenazi, Yemenite). *Textus 21*(1), 235–264.

Sacchi, P. (1973). Analisi quantitativa della tradizione medievale del testo ebraico della Bibbia secondo le collazioni del De Rossi. *Oriens Antiquus 12*, 1–13.

Wevers, J. W. (1948). A Study in the Hebrew Variants in the Books of Kings. *Zeitschrift für die Alttestamentliche Wissenschaft 61*(1), 43.