

Ajalooliste sõnakujude allikateülesest vastendamisest keelemudelite abil

Tiina Paet (EKI teadur-vanemkeelekorraldaja)

Madis Jürviste (EKI nooremteadur-leksikograaf, TÜ doktorant)

Sandra Eiche (EKI vanemkeeletehnoloog)

EKKD-III1 „Suurte keelemudelite rakendamine leksikograafias: uued võimalused ja väljakutsed“

Taust I

- SKMid ajaloolises leksikograafias
 - varasemad katsed (Jürviste, Paet, Soosaar): 17. ja 18. saj sõnaraamatutes esinevate sõnakujude vastendamine (ingl *mapping*), nt *lohrbehr* (Stahl 1637), *loorbeer* (Göseken 1660), *loorber* (nüüdiskuju): + SKMi kommentaar. Täpseim mudel Claude
- Varasem rakendus: ÕS 1918. „Eesti keele õigekirjutuse-sõnaraamatu“ kommenteeritud väljaanne

Taust II

- Laiem eesmärk: sisend sõnakujude arengu diakrooniliseks vaateks, nt visand

viinamari (tnp märksõnakuju, mille juures vastav infokirje asub):

Wihnamarri (Stahl 1637); *wihna marri* (Göseken 1660);

Wina Marri (Vestring ?1710–1730); *wina marri* (Hupel 1780);

wīna marjad (Wiedemann 1869); *viinamari* (ÕS 1925–37)...

Taust II

vein (tnp märksõnakuju, mille juures vastav infokirje asub): *Wihn* (Stahl 1637); *Wina, Marja Wina* (Gutslaff 1648); *Wihn* (Göseken 1660); *Wiin* (Vestring ?1710–1730); *Wiin* (Helle 1732); *jodaw wiin* (Hupel 1780); *wein, wiin* (Wiedemann 1869); *wein* (ÕS 1918); *vein* (ÕS 1925–37..).

viin (tnp märksõnakuju, mille juures vastav infokirje asub): *Wihn* (Stahl 1637); *Brantwein, pallatut Wina* (Gutslaff 1648); *Brantwein, Pollewiin* (Vestring ?1710–1730); *wiin* (Hupel 1780); *wīn* (Wiedemann 1869); *wiin* (ÕS 1918); *viin* (ÕS 1925–37..).

Taust III

- Alameesmärk: leida meetod struktureerimata massandmetest lemmade vastendamiseks
 - Pilotkatse eesmärk: katsetada *h*-alguliste lemmade vastendamise edukust

Näiteid sisendandmete kohta

aid, aia, -a, -ade, -u; aia|ke[ne];
aian|dus; -|ik²; (üle-) -|ne²;
-|nik

Õs
1918:
aidnik

aed, aia, 340; aedik (aiake,
sulg [-lu]), 144; ~linn,
~uba (= türgi uba),
~vili jne.; [üle-] ~ne, 166;
~nik, 496

ÕS 1940:
aednik

kael, -a, -a, -ade ja -te, -u;
-a|kas; -akute; -a|line;
-astikku; (kange-) -|ne²;
(kange-) -|sus; -|uke[ne];
-|us²; -us|tama

ÕS 1918:
kaelakute

kaela/ar'ter anat. (arteria cervicalis);
~auk = ~ava; ~ehe [-ehte];
~h'aav; ~h'aigus; ~ike etn.;
~ka'n'dja laps, kes juba «kaela
kannab»; töövõimeliseks saanud, end
ise ülalpidav nooruk; ~kas, -ka,
-kat¹⁰ tugeva kaelaga; ~k'ee;
~ke[ne], -se²⁴; ~ke't't; ~kohus
[-kohtu] van. kriminaalkohus;
~kon't; ~kuti [koos]; ~künnap;

ÕS 1976:
kaelakuti

elasti|line, kerkne

ÕS 1918:
elastiline

ÕS 1976:
elastne

ela'stik, -u³ tekst. teat. suure venivuu-
sega niit v. riie; ela'stjas, -ja, -jat¹⁰
kaunis elastne; ela'stne, -se¹⁷ vet-
ruv; ela'stselt

Pilootkatsest I

- Valim: 1925.–1937., 1940. ja 1953. a ÕS-ide *h*-algulised märksõnad
- Sisend: PDFid (pilt ilma tähtsustusega)
 - nende analüüs keerulisem: masinloetaval kujul struktureeritud andmestikega puudub parsimise vajadus
- Töötlus: parsimine > vektoriseerimine > *retrieval* ja analüüs
- Väljund: JSON, sõnakujude read allikaviidetega

Pilootkatsest II

- Väljundfailis 6198 rida (ÜSi *h*-algulised märksõnad)
 - musta materjali puhastamine:
 - ~62% ridu, millel puudub varasemates ÕSides ekvivalent, nt *hambahaldjas, hamburger, heteroabielu*
 - ~22% lemmadest duplikaadid, nt *habesamblik*
 - analüüsitav materjal: ~970 lemmat (sh vigu ja kasutuid).

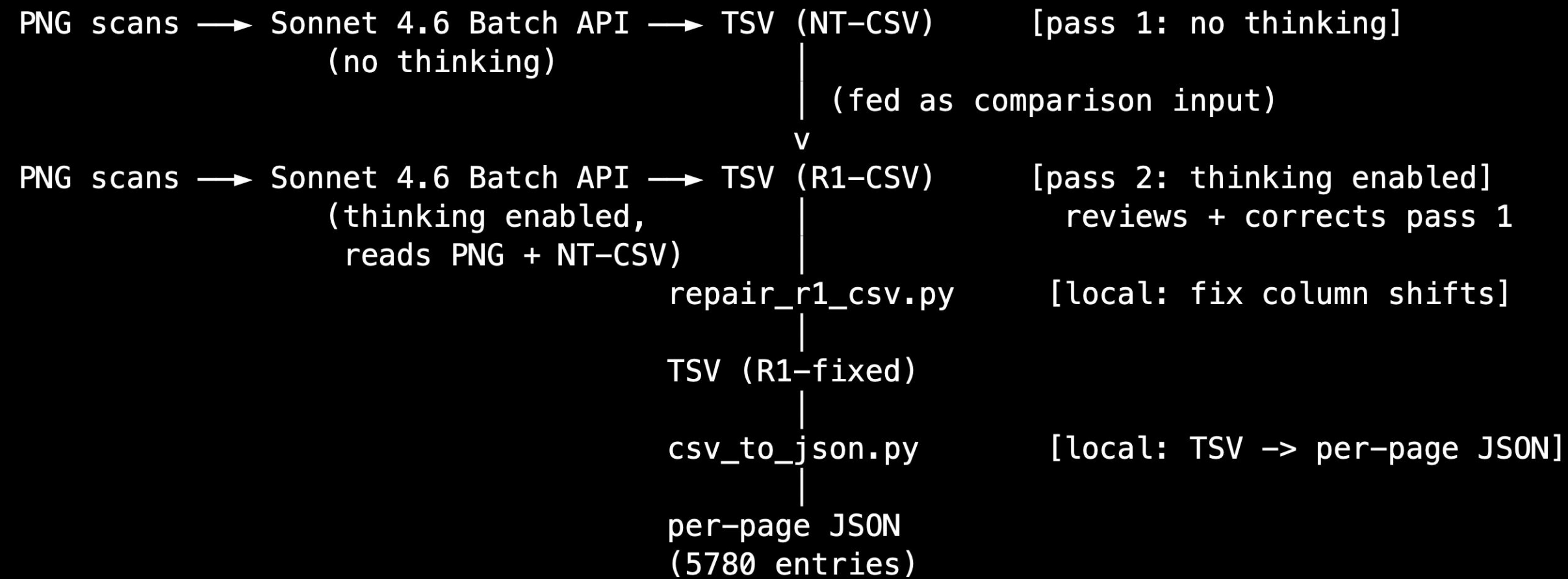
Näiteid tulemustest

- haigussümptom (ÜS) -sümptoom (ÕS 1925)
- hüdromehaanika (ÜS)
hüdromehhaanika (ÕS 1925)
- hüdroksiid (ÜS) hüdroksüüd (ÕS 1925)
- hambaplomm (ÜS) hambaplomb (ÕS 53)
- handsa (ÜS) hanža (ÕS 1925)
- harakas (ÜS) hõrakas (ÕS 1925)
- hiniin (ÜS) kiniin (ÕS 1925)
- hospital (ÜS) hospidal (ÕS 1925, 1940, 1953)
- hulkurlus (ÜS) hulkurus (ÕS 1915)
- husaar (ÜS) husar (ÕS 1925, 1940)
- huumor (ÜS) humoor (ÕS 1925)
- hüüumärk (ÜS) hüüatusemärk (ÕS 1925)
- hüppama (ÜS) hippuma (ÕS 1925)
- hüpitama (ÜS) hiputama (ÕS 1925)
- hüpoteeklaen (ÜS) hüpotekaarlaen (ÕS 1925, ÕS 1940)
- heeringakuningas (ÜS)
heeringkuningas (ÕS 1925)

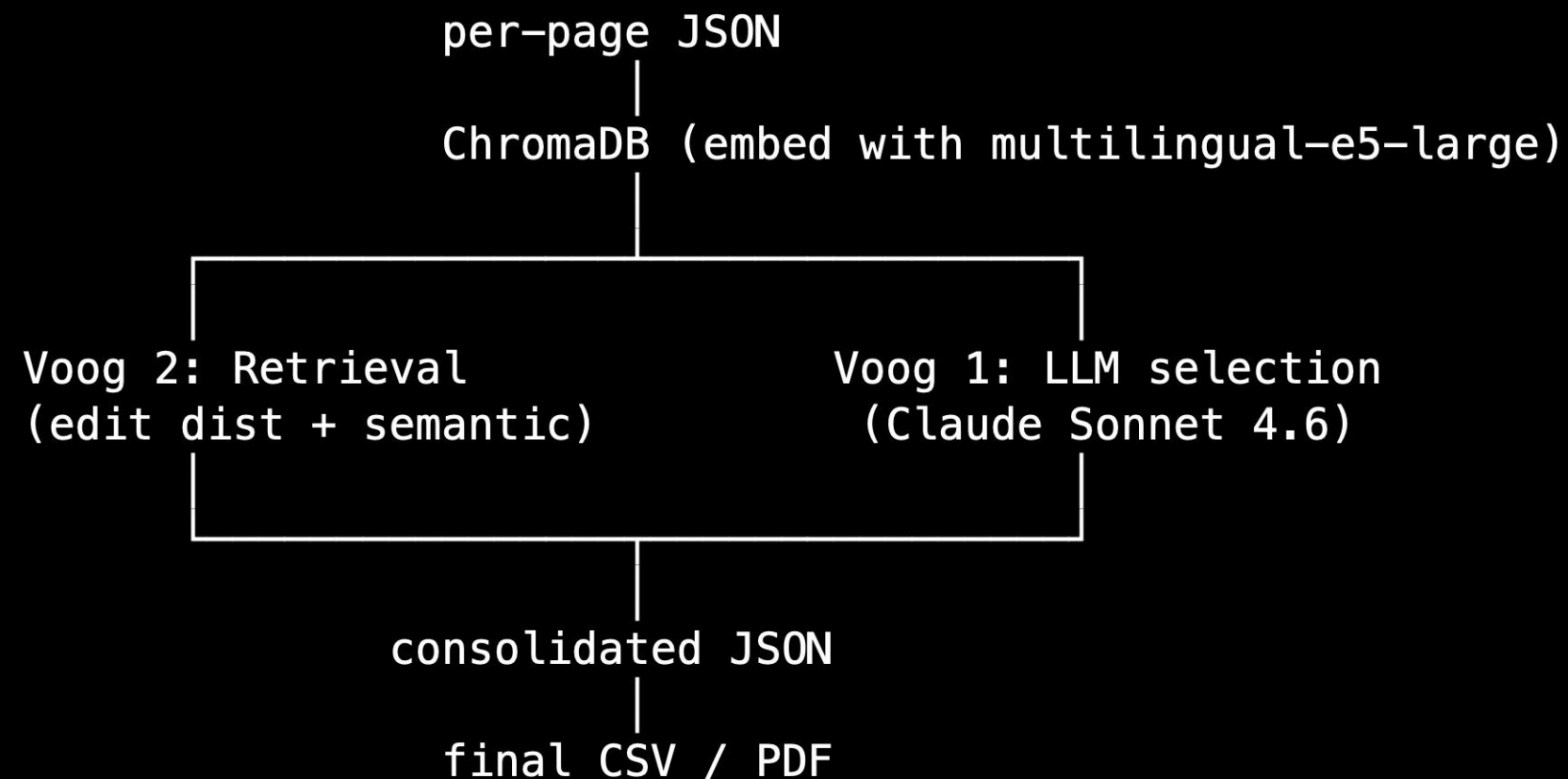
Pilootkatse tulemused

- SKMide abil on mõistliku kulu ja vaevaga võimalik süsteemselt ja massiliselt vastendada eri ajastute sõnakujusid ja tähendusi, nt *hageja* (nüüdiskuju), *hageleja* (ÕS 1925–37); *hageja* ÕS 1953
 - vead on läbipaistvad (eri etappidest: tuvastamine, seostamine), nt *hoovinaer* ja *hoovinarr*; *haavikuisand* ja *haavikuemand* (ÜS, ÕS 1925); *hambaratas* (pro *hammasratas*)
- Tulemuseks struktureeritud andmestik; eeldatud ca 5% (varaste sõnastike puhul palju suurem!) sõnakujumuutusi kogumassist on reaalne sisend diakroonilise sõnavarakihistuse loomiseks.

PHASE 1: EXTRACTION (PNG to structured CSV)



PHASE 2: MATCHING (historical entries → modern lemmas)



Töövoog

Töövoog: üldvaade

- Sisend I: PDF > PNG > base64
- Sisend II: sõnastike kasutusjuhendid
- Sisend III: prompt (vt allpool)
- Claude Sonnet 4.6: kõigi lemmade (märksõnad, allmärksõnad, loetelusõnad) lugemine pildilt 5800 + andmestiku rikastamine (ingl def, sks vaste) vektori lisaankruteks > JSON
- Vektoriseerimine, vektorotsing (ÜS 6198 ms+def vs. ÕSid), vastus

Töövoog: töötlus

- Claude Sonnet 4.6
 - Parsimine on võrdlemisi täpne (?)
 - Väljakutse: esmane CSV. Nihked. Plaasterparandused.
 - Tokenikulu säästmiseks I faasis CSV > JSON lokaalselt.
- Vektor
 - *Retrieval* on kiire, täpsus sõltub väga suurel määral filtritest:
 - POS, sem. dist., adaptiivne Levenshteini distantts jms.

haar (= haru), haara, 219;
haarak, 144; **haarakil** [ole-
ma] (haaramas); **haarakile**
[jäama]; ~**ama**, 796; **haa-
rang**, 146; ~**d_jalg** (haara-
misjalg), ~**d_juur**, ~**d_saba**
jne.; ~**duma**, 683; **haare**
(haaramine), ~**de**, 524; **haa-
re** (haaramisvahend), ~**me**,
581

```
{  
  "lemma": "haarang",  
  "lemma_source": "haarang",  
  "sõnaliik": "s",  
  "def_et": [  
    "146"  
  ],  
  "definitioon_en": [  
    "raid, sweep, roundup"  
  ],  
  "tõlkevaste_de": [  
    "Razzia, Streifzug"  
  ]  
},
```

ÕS 1940

haarama, haarata¹⁰²; **haarang**,
-u³

```
{  
  "lemma": "haarang",  
  "lemma_source": "-u",  
  "sõnaliik": "s",  
  "def_et": null,  
  "definitioon_en": [  
    "raid, sweep"  
  ],  
  "tõlkevaste_de": [  
    "Razzia, Durchsuchung"  
  ]  
},
```

ÕS 1953

**Esmane
JSON**

=====
Lemma: haarang

[1] Täpsed vasted:

ÕS 1940: haarang | raid, sweep, roundup | ET: 146 | DE: Razzia, Streifzug

ÕS 1953: haarang | raid, sweep | DE: Razzia, Durchsuchung

[3] Semantiline otsing ankrutega (14 tulemust):

[1940] edit=0 sem=0.0323 | haarang: haarang | raid, sweep, roundup | ET: 146 | DE: Razzia, Streifzug

[1953] edit=0 sem=0.0626 | haarang: haarang | raid, sweep | DE: Razzia, Durchsuchung

[1925] edit=8 sem=0.1054 | huligaansus: huligaansus | hooliganism | DE: Rowdytum, Hooliganismus

[1953] edit=3 sem=0.1058 | harali: harali | spread out, apart | DE: auseinander, gespreizt

[1940] edit=6 sem=0.1096 | hoiatus: hoiatus | warning, caution | DE: Warnung

[1925] edit=4 sem=0.1111 | häving: häving | ruin, destruction | DE: Vernichtung, Untergang

[1925] edit=5 sem=0.1275 | heeringas: heeringas | herring | DE: Hering

[1940] edit=4 sem=0.1286 | haardsaba: haardsaba | prehensile tail | DE: Greifschwanz

[1925] edit=8 sem=0.1300 | heeringapüük: heeringapüük | herring fishing | DE: Heringsfang

[1953] edit=8 sem=0.1309 | heeringapüük: heeringapüük | herring fishing | ET: heeringapüük | DE: Heringsfang

[1925] edit=6 sem=0.1318 | heinareha: heinareha | hay rake | DE: Heurechen

[1925] edit=6 sem=0.1318 | heinasaak: heinasaak | hay crop | DE: Heuertrag

[1940] edit=2 sem=0.1320 | haarak: haarak | fork, branch | ET: 144 | DE: Gabel, Ast

[1925] edit=8 sem=0.1326 | heeringatünn: heeringatünn | herring barrel | DE: Heringsfass

PASS (edit≤2): [1940] edit=0 sem=0.0323 | haarang

PASS (edit≤2): [1953] edit=0 sem=0.0626 | haarang

PASS (edit≤2): [1940] edit=2 sem=0.1320 | haarak

[4] Lõplikud kandidaadid LLM-ile:

ÕS 1925: (tühi)

ÕS 1940: haarang: haarang | raid, sweep, roundup | ET: 146 | DE: Razzia, Streifzug

ÕS 1940: haarak: haarak | fork, branch | ET: 144 | DE: Gabel, Ast

ÕS 1953: haarang: haarang | raid, sweep | DE: Razzia, Durchsuchung

```
{
  "input_lemma": "haarang",
  "input_definition": "mingi maa-ala sissepiiramine ja läbiotsimine jälitatava(te)
tabamiseks; RU: облава, гонение, травля",
  "matches": {
    "1925": [],
    "1940": [
      {
        "lemma": "haarang",
        "year": 1940,
        "page": 1,
        "doc": "haarang | raid, sweep, roundup | ET: 146 | DE: Razzia, Streifzug"
      }
    ],
    "1953": [
      {
        "lemma": "haarang",
        "year": 1953,
        "page": 1,
        "doc": "haarang | raid, sweep | DE: Razzia, Durchsuchung"
      }
    ]
  }
},
{
```

Lõppväljundi JSON

Tulemused

- Väljakutse ja piirangud
 - käsitsianalüüsiga võrreldes märkimisväärne ajasääst.
Tokenikulu (sisend+väljund) ca 80 €.
- Tulemus
 - Töövoo ülesehitus (PDF > tabel vastendatud lemmadega)
 - kasutatav ka vanade sõnastike puhul
 - Perspektiivid: kõik ÕSid vastendatud, seejärel laiendus muudele sõnastikele

Kirjandust

Jürviste, Madis; Paet, Tiina; Soosaar, Sven-Erik 2025. Eesti vanade sõnakujude tuvastamise võimalustest suurte keelemudelite abil. – Eesti Rakenduslingvistika Ühingu aastaraamat 21.

Paet, Tiina 2026. Kuidas *viinast* sai *vein*: sõnastiku märksõna diakroonilise kirjelduskihi visand. Keel ja Kirjandus (käsikiri valmimas)

AITÄH

Prompt (*incipit*)

Roll: Sa oled leksikograaf ja korpuse-/andmetöötlaste spetsialist, kes struktureerib vanade õigekeelsussõnaraamatute (QS) sõnaartikleid masintöödeldavaks andmestikuks.

Sihtrühm: doktorikraadiga keeleteadlased.

0) Sisend

Sul on alati kaks sisendit:

** PNG-failid (üks või mitu järjestikust lehekülge), mis sisaldavad sõnaartikleid (nt ühe tähe alajaotus, siin H).*

** Kasutusjuhised (md) vastava QS-i jaoks (võib erineda aastakäiguti). Kasutusjuhised on normatiivne allikas, mille järgi tõlgendada kõiki märke, eraldajaid ja lühendeid.*

...

>>> [Prompti pikkus kokku 6855 tm e ca 4 lk.] <<<