Eeltöötlusmooduli dokumentatsioon

Eesmärk: süntaksianalüüsi-eelse eeltöötlusmoodul parandab korpustes lausestusvigu ja sagedasemaid süntaktilise analüüsi vigu põhjustavad vead. Lausestusvigadest parandatatakse selliseid juhtumeid, kus lausemärgendid on üleliigsed. Süntaktilise analüüsi vigu põhjustavate vigade korral ühendati lause sees paiknevad mitteleksikaalsed elemendid täiendava märgendiga (<+/>). Spordiuudiste ja saatekava lõikude puhul ignoreeritakse mitte-lauselist sisu.

1. Arvud ja numbrid:
   - 20 000 ⇒ 20<+/>000
   - 669 81 54 ⇒ 669<+/>81<+/>54
   - 1 , 2 ⇒ 1,2
   - 25 - 30 % ⇒ 25-30% , 20 %ga ⇒ 20%ga, 0,20 -protsendilise ⇒ 0,20-protsendilise
   - lühend (nt AS) või jutumärk, kaldkriips, § või arv, millele järgneb käändelõpp:
     - 1,5 -ni ⇒ 1,5-ni
     - § -st ⇒ §-st
   - avaldised:
     - 24+9 = 33 ⇒ 24+9<+/>=<+/>33
     - R 2 = 0 , 08 ⇒ R<+/>2<+/>=<+/>0,08
     - r = 0 , 46 ⇒ r<+/>=<+/>0,46
   - arv, millele järgneb mõõtühiku lühend:
     - 2 t 15 min ⇒ 2<+/>t 15<+/>min
     - 294 ha-lt ⇒ 294<+/>ha-lt
     - 1 ha ⇒ 1<+/>ha
     - 1,0 mM ⇒ 1,0<+/>mM
     - 740 kHz-ni ⇒ 740<+/>kHz-ni
     - 7 C°⇒ 7<+/>C°
     - 60 km / h ⇒ 60<+/>km/h
     - 2,3 h / m –> 2,3<+/>h/m
     - 443,49 kr/MWh ⇒ 443,49<+/>kr/MWh
     - 22 000 kr/m^2 ⇒ 22<+/>000<+/>kr/m^2
   - aastarvud ja kuupäevad:
     - 21. </s> <s> 12. 2001 ⇒ 21.12.2001
     - 1884. a. ⇒ 1884.a.
     - 1884 a. ⇒ 1884a.
     - 15. 04. 2005 ⇒ 15.04.2005
     - 1945 . aasta ⇒ 1945. aasta
2. Vahemikud:
   - 40 000-45 000 ⇒ 40<+/>000<+/>-<+/>45<+/>000
   - 0 , 3 . . . </s> <s> 1% ⇒ 0,3<+/>...<+/>1%

- 1998. - 2000 ⇒ 1998.<+/>-<+/>2000
- 2,0 ... </s> <s> 3,5 ⇒ 2,0<+/>...<+/>3,5
- 40 − 300 C°⇒ 40<+/>−<+/>300<+/>C°
- 3 ... 8 mÜs ⇒ 3<+/>...<+/>8<+/>mÜs
- 4 . − 5 . ⇒ 4.<+/>−<+/>5.

3. Nimed:
   - J. Fr . </s> <s> Blumenbach ⇒ J.Fr.<+/>Blumenbach
   - A . </s> <s> J. Sjögren ⇒ A.J.<+/>Sjögren
   - J. R. R. Tolkieni ⇒ J.R.R.<+/>Tolkieni
   - St. Louis ⇒ St.<+/>Louis
   - Simon & Schusteri ⇒ Simon<+/>&<+/>Schusteri
   - Laulasmaa Spa & Konverentsihotell ⇒ Laulasmaa Spa<+/>&<+/> Konverentsihotell
4. Lausestusvead:
   - (ala)jaotused, loetelud, järgarvud:
     - (2 . </s> <s> 8) ⇒ (2 . 8)
     - (85 . </s> <s> Antonov) ⇒ (85 . Antonov)
     - 1 . </s> <s> 4. 3. </s> <s> Tahtlus /.../ ⇒ 1.4.3. Tahtlus /.../
     - 5 . </s> <s> 2. 3. </s> <s> Ujumisõpetajad ⇒ 5.2.3.Ujumisõpetajad
   - veebiaadressid:
     - www. </s> <s> RODesign.ee/ ⇒ www.RODesign.ee/
   - lauselõpud, kus punkt on stringi külge jäänud:
     - ajal. </s> Peame /.../ ⇒ ajal . </s> Peame /.../
     - jne. </s> <s> Aga udmurdid /.../ ⇒ jne . </s> <s> Aga udmurdid /.../
     - aastasse 2000. </s> <s> Ja teine /.../ ⇒ aastasse 2000 . </s> <s> Ja teine /.../
5. Spordiuudised:
   - tulemused, ajad:
     - 7 : 8 ⇒ 7<+/>:<+/>8
     - 2 . 06 , 08 ⇒ 2<+/>.<+/>06<+/>,<+/>08
6. <**ignore**> ja </**ignore**> märgendid:
   - sulgudes olevad arvud ja/või lühendid, suurtähega algavad sõned:
     - (WTA 210.) ⇒ <**ignore**> (WTA 210.) </**ignore**>
     - Kreekaga (57.) ⇒ Kreekaga <**ignore**> (57.) </**ignore**>
     - Tuneesia (5) ⇒ Tuneesia <**ignore**> (5) </**ignore**>
   - pikad loetelud:

- <s> Rahvamajandus : Abiševa , Maria ; Ahlamtšenkova , Viktoria ; /.../ Vinkel , Natalja ; Zahharov , Sergei . </s> ⇒ **<ignore>** <s> Rahvamajandus : Abiševa , Maria ; Ahlamtšenkova , Viktoria ; Bakulina , Maria ; /.../ Vinkel , Natalja ; Zahharov , Sergei . </s> **</ignore>**
- <s> Väravad : Vahtramäe (5) , Ustritski (7) , M. Rooba (34) - Voronin (30) , Leitan (42) </s> ⇒ **<ignore>** <s> Väravad : Vahtramäe (5) , Ustritski (7) , M.<+/>Rooba (34) - Voronin (30) , Leitan (42) </s> **</ignore>**
- <s> Tabeliseis : Austria 6 punkti , Poola 4 , Leedu ja Holland 3 , Eesti 2 , Horvaatia 0 . </s> ⇒ **<ignore>** <s> Tabeliseis : Austria 6 punkti , Poola 4 , Leedu ja Holland 3 , Eesti 2 , Horvaatia 0 . </s> **</ignore>**

- spordiuudised:
  - lõigud või laused, mis algavad spordiala ja/või distantsiga ning millele järgneb tulemuste loetelu:
    * <s> 5 km (v) : 1. Katerina Neumannova (pildil , Tshehhimaa) 12.56 , 1 , 2. </s> <s> Tshepalova +9,0 , 3. </s> /.../ <s> Jelena Buruhina (Venemaa) +39,2 , 28. </s> <s> Skari +53,8 , ... </s> <s> 65. Katrin *migun +1.33 , 0. </s> ⇒ **<ignore>** <s> 5 km (v) : 1. Katerina Neumannova (pildil , Tshehhimaa) 12.56 , 1 , 2. </s> <s> Tshepalova +9,0 , 3. </s> /.../ <s> Jelena Buruhina (Venemaa) +39,2 , 28. </s> <s> Skari +53,8 , ... </s> <s> 65. Katrin *migun +1.33 , 0. </s> **</ignore>**
    * <s> Sprint : 1. Ronny Ackermann (Saksamaa) 18.58 , 2 , 2. </s> <s> Hannu Manninen (Soome) +9,3 , 3. </s> <s> Kristian Hammer (Norra) +10,4 , 4. </s> /.../ <s> Christoph Bieler (Austria) +1.09 , 1 , 25. </s> <s> Jens Salumäe (Eesti) +1.31 , 0. </s> ⇒ **<ignore>** <s> Sprint : 1. Ronny Ackermann (Saksamaa) 18.58 , 2 , 2. </s> <s> Hannu Manninen (Soome) +9,3 , 3. </s> <s> Kristian Hammer (Norra) +10,4 , 4. </s> /.../ <s> Christoph Bieler (Austria) +1.09 , 1 , 25. </s> <s> Jens Salumäe (Eesti) +1.31 , 0. </s> **</ignore>**
    * <s> Tulemused . </s> <s> Naiste 1000 m : 1 . </s> <s> Chris Witty (USA) 1 . 14 , 96 , 2 . </s> /.../ <s> Elena Belci (Itaalia) 7 . 16 , . </s> ⇒ <s> Tulemused . </s> **<ignore>** <s> Naiste 1000<+/>m : 1 . </s> <s> Chris Witty (USA) 1<+/>.<+/>14<+/>,<+/>96 , 2 . </s> /.../ <s> Elena Belci (Itaalia) 7<+/>.<+/>16<+/>,<+/>43 . </s> **</ignore>**
    * <s> Mehed : 200 m 1 . </s> <s> Maurice Greene (USA) 19 , 92 , 2 . </s> <s> Frank Fredericks (Namibia) 19 , 93 , 3 . </s> /.../ <s> Younes Moudrik (Maroko) 8 . 20 , 3 . </s> <s> Kareem Streete-Thompson (Kaimanisaared) 8 . 15 . </s> ⇒ **<ignore>** <s> Mehed : 200<+/>m 1 . </s> <s> Maurice Greene (USA) 19 , 92 , 2 . </s> <s> Frank Fredericks (Namibia) 19 , 93 , 3 . </s> /.../ <s> Younes Moudrik (Maroko) 8<+/>.<+/>20<+/>,<+/>3 . </s> <s> Kareem Streete-Thompson (Kaimanisaared) 8 . 15 . </s> **</ignore>**
    * <s> <hi rend="rasvane"> 10 000 m : </hi> 1 . </s> <s> Sally Barsosio (Keenia) 31 . 32 , 29 , 2 . </s> /.../ <s> Chemi Takahashi (Jaapan) 32 . 23 , >61 . </s> ⇒ **<ignore>** <s> <hi rend="rasvane"> 10<+/>000<+/>m : </hi> 1 . </s> <s> Sally Barsosio (Keenia) 31<+/>.<+/>32<+/>, <+/>29 , 2 . </s> /.../ <s> Chemi

Takahashi (Jaapan) 32<+/>.<+/>23<+/>,<+/>61 . </s> </**ignore**> </div4>

  * <s> <hi rend=“rasvane”> 400 m : </hi> 1 . </s> <s> Iwan Thomas (Suurbritannia) 44 , 90 , 2 . </s> /.../ <s> Chris Jones (USA) 45 , 71 , 6 . </s> <s> Sunday Bada (Nigeeria) 45 , 86 . </s> ⇒ <**ignore**> <s> <hi rend=“rasvane”> 400<+/>m : </hi> 1 . </s> <s> Iwan Thomas (Suurbritannia) 44 , 90 , 2 . </s> /.../ <s> Chris Jones (USA) 45 , 71 , 6 . </s> <s> Sunday Bada (Nigeeria) 45 , 86 . </s> </**ignore**>
  * <s> <hi rend=“rasvane”> Maraton </hi> 2.25.17 Rosa Mota , Portugal 1987 </s> ⇒ <**ignore**> <s> <hi rend=“rasvane”> Maraton </hi> 2.25.17 Rosa Mota , Portugal 1987 </s> </**ignore**>

– meeskonna või sportlase nimi, millele järgnevad tulemused (numbrijada):

  * <s> Joe Sakic Col 32 16+27= 43 10 </s> ⇒ <**ignore**> <s> Joe Sakic Col 32<+/>16+27=<+/>43<+/>10 </s> </**ignore**>
  * <s> NY Islanders 20 6 3 0 11 15 </s> ⇒ <**ignore**> <s> NY Islanders 20<+/>6<+/>3<+/>0 11<+/>15 </s> </**ignore**>
  * <s> Miami-Orlando 2 : 2 </s> ⇒ <**ignore**> <s> Miami-Orlando 2<+/>:<+/>2 </s> </**ignore**>
  * <s> NEW JERSEY 14 29 . </s> <s> 326 18 1/2 </s> ⇒ <**ignore**> <s> NEW JERSEY 14<+/>29 . </s> <s> 326<+/>18<+/>1/2 </s> </**ignore**>
  * <s> Zimbru Chisinau (Moldova) – Dinamo Tbilisi (Gruusia) 2 : 0 (kokku 3 : 2) . </s> /.../ <s> Dariusz Gesior 14 , Artur Wichniarek 52 , 60 , Radoslaw Michalski 75 – Svetoslav Todorov 30 . </s> ⇒ <**ignore**> <s> Zimbru Chisinau (Moldova) – Dinamo Tbilisi (Gruusia) 2<+/>:<+/>0 (kokku 3<+/>:<+/>2) . </s> /.../ <s> Dariusz Gesior 14 , Artur Wichniarek 52 , 60 , Radoslaw Michalski 75 – Svetoslav Todorov 30 . </s> </**ignore**>
  * <s> Teised kohtumised : Washington - Colorado 2 : 1 , Ottawa - Carolina 5 : 1 , Tampa Bay - NY Rangers 2 : 2 (la) , NY Islanders - San Jose 1 : 4 , St . </s> <s> Louis - Atlanta 4 : 1 , Nashville - Edmonton 1 : >2 . </s> ⇒ <**ignore**> <s> Teised kohtumised : Washington - Colorado 2<+/>:<+/>1 , Ottawa - Carolina 5<+/>:<+/>1 , Tampa Bay - NY Rangers 2<+/>:<+/>2 (la) , NY Islanders - San Jose 1<+/>:<+/>4 , St . </s> <s> Louis - Atlanta 4<+/>:<+/>1 , Nashville - Edmonton 1<+/>:<+/>2 . </s> </**ignore**>
  * <s> 2. Svetlana Tšernoussova Venemaa +1.12 , 8 (1) </s> ⇒ <**ignore**> <s> 2. Svetlana Tšernoussova Venemaa +1.12 , 8 (1) </s> </**ignore**>
  * <s> 1. Valencia 17 10 5 2 29 : 10 35 </s> ⇒ <**ignore**> <s> 1. Valencia 17<+/>10<+/>5<+/>229<+/>:<+/>10<+/>35 </s> </**ignore**>
  * <s> K Clijsters (BEL) (15) - A Jidkova (RUS) 6 - 3 , 7 - 6 </s> ⇒ <**ignore**> <s> K Clijsters (BEL) (15) - A Jidkova (RUS) 6<+/>-<+/>3 , 7<+/>-<+/>6 </s> </**ignore**>

- viited teadustekstides (vt PATT_BRACS, PATT_12):

  - (vt joonis 4.6) ⇒ <**ignore**> (vt joonis 4.6) </**ignore**>
  - (2001 : 114) ⇒ <**ignore**> (2001 : 114) </**ignore**>
  - (Silver , 1992 ; Alessi& Trollip , 2001 : 115) ⇒ <**ignore**> (Silver , 1992 ; Alessi& Trollip , 2001 : 115) </**ignore**>
  - (1984) ⇒ <**ignore**> (1984) </**ignore**>
  - (Miljan , R. et al. Choices for Dairy ... 2003) ⇒ <**ignore**> (Miljan , R. et al. Choices for Dairy ... 2003) </**ignore**>
  - (tabel 3.2) ⇒ <**ignore**> (tabel 3.2) </**ignore**>
- sulgude sees paiknevad lühendid, sõnede ja arvude kombinatsioonid (vt PATT_BRACS, PATT_12):
  - (A – 23,87 tuhat krooni , B – 23,72 tuhat krooni ja D – 23,35 tuhat krooni) ⇒ <**ignore**> (A – 23,87 tuhat krooni , B – 23,72 tuhat krooni ja D – 23,35 tuhat krooni) </**ignore**>
  - (195 miljonit USDd) ⇒ ignore> (195 miljonit USDd) </**ignore**>
  - (à 30 min) ⇒ <**ignore**> (à 30<+/>min) </**ignore**>
  - (“ Hannah ja ta õed ” , 1986) ⇒ <**ignore**> (“ Hannah ja ta õed ” , 1986) </**ignore**>
  - (Hispaania , 3) ⇒ <**ignore**> (Hispaania , 3) </**ignore**>
  - (rasvasus - 12% , valgusisaldus - 16%) ⇒ <**ignore**> (rasvasus - 12% , valgusisaldus - 16%) </**ignore**>
  - (2Kr 6,10) ⇒ <**ignore**> (2Kr 6,10) </**ignore**>
  - (15 juhtumit , kusjuures 2 said surma) ⇒ <**ignore**> (15 juhtumit , kusjuures 2 said surma) </**ignore**>
  - (50 : 47,81 : 81 ; Jamchy 20 , Henefeld 19 - Gadou 25 , Rigaudeau 24) ⇒ <**ignore**> (50 : 47,81 : 81 ; Jamchy 20 , Henefeld 1 9<+/>-<+/>Gadou 25 , Rigaudeau 24) </**ignore**>
  - (RKO) ⇒ <**ignore**> (RKO) </**ignore**>
  - (viimasel päeval võitis Goran Ivanisevic (Horvaatia) Thomas Musteri 6 : 7 , 7 : 5 , 6 : 7 , 6 : 2 , 7 : 5) , Zimbabwe-Suurbritannia 4 : 1 . </s> ⇒ <**ignore**> (viimasel päeval võitis Goran Ivanisevic (Horvaatia) Thomas Musteri 6<+/>:<+/>7,7<+/>:<+/>5,6<+/> :<+/>7,6<+/>:<+/>2,7<+/>:<+/>5) </**ignore**> , Zimbabwe-Suurbritannia 4<+/>:<+/>1 . </s>
  - (100 ja 4x100 m) ⇒ <**ignore**> (100 ja 4x100 m) </**ignore**>
  - <s> 26-aastase Aleksander Tammerti seeria (65.36 - 59.44 - 66.95 - 62.06 - 62.50 - 66.48) andis /.../ ⇒ <s> 26-aastase Aleksander Tammerti seeria <**ignore**> (65.36-59.44-66.95-62.06-62.50-66.48) </ignore> andis /.../
- telekava:
  - <s> <hi rend=“rasvane”> 06 . 00 </hi> Saatekavateade <hi rend=“rasvane”> 06 . 05 </hi> /.../ Õigel ajal <hi rend=“rasvane”> 23 . 05 </hi> Venemaa aidsi vastu <hi rend=“rasvane”> 23 . 55 </hi> Ilmateade <hi rend=“rasvane”> 00 . 00 </hi> Kunstipoodium <hi rend=“rasvane”> 00 . 35 </hi> Telepood </s> ⇒ <**ignore**> <s> <hi rend=“rasvane”> 06 . 00 </hi> Saatekavateade <hi rend=“rasvane”> 06 . 05 </hi> /.../ Õigel ajal <hi rend=“rasvane”> 23 . 05 </hi> Venemaa aidsi vastu <hi rend=“rasvane”> 23 . 55 </hi> Ilmateade <hi rend=“rasvane”> 00 . 00 </hi> Kunstipoodium <hi rend=“rasvane”> 00 . 35 </hi> Telepood </s> </**ignore**>

– <s> <hi rend="rasvane"> 07 . 00 </hi> Tere hommikust ! </s> ⇒ **<ignore>** <s> <hi rend="rasvane"> 07 . 00 </hi> Tere hommikust ! </s> **</ignore>**

7. Igale lausele lisati selle järjekorranumber tekstis (va **<ignore>** ja **</ignore>** märgendite vahel olevatele lausetele):

   - <s> <id="15">
   - </s> <s> <id="1066">