Lauma Pretkalniņa, Laura Rituma
Mākslīgā intelekta laboratorija,
LU MII

Latviešu valodas sintaktiski marķētā korpusa teksta sadalīšanas tekstvienībās un morfoloģiskās marķēšanas metodikas apraksts

1. Teksta sadalīšana tekstvienībās

Teksta sadalīšana tekstvienībās notiek pamatā automātiski, bet dažos gadījumos rīka piedāvātais rezultāts vēl tiek mainīts manuāli – vairākas tekstvienības tiek apvienotas vienā vai viena tekstvienība tiek sadalīta divās. Šo manuālo iejaukšanos var redzēt arī morfoloģijas līmenī, jo izmaiņas tiek atzīmētas .m failos, norādot izmaiņu iemeslu un veidu, kā arī jaunās tekstvienības morfoloģisko tagu. Vizuāli manuālas tekstvienību robežu izmaiņas var redzēt TRED rīkā – teikuma attēlošanas logā tās tiek iekrāsotas.

1.1. Tekstvienības, kuras veido vairāki teksta elementi

Kā viena tekstvienība tiek marķēti saīsinājumi ar pieturzīmēm pa vidu:

- saīsinājumi un iniciāļi ar punktu beigās vai arī pa vidu, piem., **utt.**, **u.tml.**, **u.c.**, **utjp.**, **funkc.**, **A.**;
- kārtas skaitļi, kas atveidoti ar cipariem un beidzas ar punktu, kā arī datumi, piem., 13., 21.09.1976.;
- decimāldaļskaitļi, kuriem pa vidu ir komats, piem., 14,57;
- pulksteņlaiki, kuros izmanto punktu vai kolu, piem., 15:00, 15.00, 15:34:02.

Kā viena tekstvienība tiek marķēti defissavienojumi:

- apzīmējumi, kuros defise ir lietota starp vārda daļām un liecina par vārda īsinājumu, piem., **k-gs**;
- defissavienojumi, kuros pirmais komponents ir iniciālisms – motivētājvārda pirmais burts, bet otrais komponents ir pilns vārds, piem., **e-klase, e-talons, e-paraksts, i-banka, i-veikals**;
- izsauksmes vārdi, kuri sastāv no defissavienojuma, piem., **hi-hi, ku-kū, vau-vau**;
- vairākvārdu uzvārdi, kuros īpašvārdi savienoti ar defisi, piem., **Vīķe-Freiberga**.

Viena tekstvienība ir arī tādi saīsinājumi, starp kuru daļām iespējama atstarpe:

- skaitļi un telefona numuri (xn), piemēram, *2 358 000* un *+371 20001234*;
- saīsinājumi *u.t.jpr., u.c., u.tml., v.tml.*, *u.t.t.*, *N.B.*, *P.S.* un *P.P.S.* ar jebkuru *P.* daudzumu, ja tajos ir atstarpes.

1.2. Teksta elementi, kas sadalāmi vairākās tekstvienībās

Dažos gadījumos automātiskā teksta sadalīšana tekstvienībās ir izveidojusi vienu tekstvienību, kuru tomēr nepieciešams manuāli sadalīt vairākās:

- defissavienojumi, kuros ar defisi savienoti divi patstāvīgi vārdi (parasti sugasvārdi), piemēram, **rīsu-griķu**, **lāzeriem-mērierīcēm, Rietumu-Austrumu**.

1.3. Tekstvienību labojumu atveide morfoloģijas slānī (.m failos)

Tekstvienības tiek labotas šādos gadījumos:

1) ja oriģināltekstā ir atsevišķu burtu kļūdas, piemēram, "sķiet" → "šķiet",
2) ja vārds oriģināltekstā kļūdas dēļ ir uzrakstīts vairākos vārdos, piemēram, *tas ir ne vajadzīgs*;
3) ja oriģināltekstā bijis sarakstīts kopā kaut kas, ko mēs gribam atdalīt, piemēram "kautkur";
4) ja oriģināltekstā ir bijis ievietots lieks komats;
5) ja oriģināltekstā trūkst komata.

Virsotnēm, kurām **atbilst kāda tekstvienība** (vārds vai pieturzīme), ir vēl citi lauki – tām ir viens **m** lauku bloks (apraksta morfoloģiju) un patvaļīgs skaits **w** bloku (katrs bloks apraksta vienu tekstvienību – piemēram vārdam „**3,14**" var būt trīs **w** bloki).

- **m/w/token** – tekstā lietotā vārdforma
- **m/w/no_space_after** – 1, ja starp doto tekstvienību un nākamo netiek lietota atstarpe (piemēram, ja šis ir „**(**" vai nākamais ir „**.**").
- **m/form** – pareizā vārdforma. Šis lauks var atšķirties no **m/w/token**, ja tekstā ir bijusi drukas kļūda vai ja vārds sastāv no vairākām tekstvienībām (piemēram, ja **m/w/token** ir **zabaks**, tad **m/form** ir **zābaks**) (obligāts).
- **m/form_change** – ja **m/form** atšķirās no **m/w/token**, šeit tiek norādīts veiktā labojuma veids. (TRED logā vietā, kur ir redzams teikums, parādās dzeltenā krāsā iekrāsots ievietots vai labots elements, izņemot gadījumu, kad laboti kopā sarakstīti vārdi (kopā sarakstīts "kaut kas")). Iespējamās vērtības (jāliek visas, kas atbilst):
    - **spell** – izlabota drukas kļūda;
    - **punct** – izlabota pieturzīmes kļūda;
    - **insert** – ielikta jauna tekstvienība, kas tekstā bija izlaista;
    - **spacing** – (a) aiz šī vārda oriģināltekstā trūkst vajadzīgas atstarpes, (b) vairākas ar atstarpi uzrakstītās tekstvienībās ir apvienotas vienā;
    - **union** – vārds izveidots no divām vai vairākām tekstvienībām;

- **num_normalization** – mainīts skaitļu pieraksts, piemēram, no AM/PM uz Latvijā lietoto, utt.

Piemēri:

- ielikts trūkstošs komats – **punct** un **insert**;
- izdzēsts lieks komats, to apvienojot ar iepriekšējo vārdu – **punct** un **union**;
- vārdam pielikta trūkstoša garumzīme – **spell**;
- atsevišķi rakstāmi vārdi (“**kaut kas”)** sarakstīti kopā (“**kautkas**”) – **spacing** ir obligāts pie pirmajām daļām, taču uz 2024-10-31 tas biežāk ir salikts pie abām;
- kopā rakstāms vārds (“**jādara**”) uzrakstīts atsevišķi (“**jā dara”**) – **spacing** un **union**;
- nepareizi sadalīts tekstvienībās, bet ar atstarpēm viss ir kārtībā (**“nozare.lv**” ir rakstīts bez atstarpēm, bet sadalīts trīs **w** blokos – “**nozare**”, “**.**”, “**lv**”) – tikai **union**.

NB! Ja tiek atdalīts defissavienojums, piemēram, “prasību-pārbaužu”, “pieņemšanas-nodošanas”, tad **m/form_change** lauks nav jāaizpilda.

NB! Ja tiek atdalīta pieturzīme no skaitļa, tad **m/form_change** lauks nav jāaizpilda.

NB! UD vienkāršības labad vienīgā vieta korpusā (d197, 100 gramu), kur ar cipariem pierakstītā skaitlī ir nepareiza atstarpe (**1 00** nevis **100**), ir nomarķēta kā **spell** – pareizrakstības kļūda.

2. Morfoloģija

Dokumentā SemTi-*Kamols_morphotags.ods* aprakstīta LU MII Mākslīgā intelekta laboratorijas (AiLab.lv) izstrādātā un teksta korpusu morfoloģiskajā marķēšanā izmantotā pazīmju kopa, kurā uzskaitītas vārdšķiras un citas tekstvienību grupas, katrai no tām norādot morfoloģisko pazīmju un to apzīmējumu (tagu) kopumu, katras pazīmes vērtību ar skaidrojumu un piemēriem.

2.1. Lietvārds

Lietvārda pamatforma parasti ir vsk. nominatīvs: ar mazo sākumburtu sugasvārdiem, ar lielo – īpašvārdiem. Izņēmumi: daudzskaitliniekam – dsk. nominatīvs (*bikses*), ģenitīvenim – ģenitīva forma (*vienstāva, augstpapēžu*).

4. un 5. deklinācijas kopdzimtes lietvārdi (*paziņa*) atkarībā no konteksta marķēti kā vīriešu dzimtes (*paziņam*) vai sieviešu dzimtes (*paziņai*) lietvārdi.

Nelokāmajiem lietvārdiem nepiemīt šādas gramatiskās kategorijas: skaitlis, locījums, deklinācija. Dzimte nelokāmajiem lietvārdiem tiek piemērota tikai tad, ja tā norādīta vārdnīcā tezaurs.lv (*kanoe* – siev. dz.).

Atšķirībā no tradicionālās gramatikas morfoloģisko pazīmju kopā nav lietots instrumentālis, jo tas vienskaitlī sakrīt ar akuzatīvu, bet daudzskaitlī – ar datīvu.

2.2. Darbības vārds

Darbības vārda pamatforma ir nenoteiksme. Noliegtas formas (*neskrien*) pamatforma ir nenoliegtā formā (*skriet*).

Darbības vārda tips tiek noteikts atkarībā no tā, vai tas teikumā lietots patstāvīgā nozīmē vai palīgnozīmē. Palīgnozīmē lietoto darbības vārdu iedalījums:

- semantiskie modificētāji (modāli, fāzes un izpausmes veida);
- palīgverbs *būt* analītiskās darbības vārda formās vai sastata izteicējā un palīgverbi *tikt*, *tapt*;
- *kļūt*, *tapt* un citi saitiņverbi (saitiņas funkcijā lietoti darbības vārdi).

Darbības vārda divdabja formas (4. pozīcijas vērtība – p) aprakstītas atsevišķā tabulā "Darbības vārda divdabja forma". To tags pirmās četras pozīcijas veidojas tādas pašas kā darbības vārdam, no kura divdabis veidots, bet 5. – 13.pozīcija ir no savas tabulas. Darbības vārda divdabja formas pamatforma ir darbības vārda nenoteiksme. Ja tekstā lietota noliegta darbības vārda divdabja forma (*neskatīdamās*), tās pamatforma ir nenoteiksme nenoliegtā formā (*skatīties*).

2.3. Īpašības vārds

Lokāmo īpašības vārdu pamatforma parasti ir vīriešu dzimtes vsk. nominatīvs pamata pakāpē, ar nenoteikto galotni (*zaļš, glīts, neglīts*). Izņēmumi: 1) adjektīviskiem uzvārdiem tiek saglabāts lielais sākumburts, dzimte, pakāpe un noteiktība, kas ir uzvārdā (*Baltā*); 2) īpašības vārdiem, kuriem nav nenoteiktās galotnes, pamatforma ir ar noteikto galotni (*pēdējais, pārējais, galvenais*).

Nelokāmo īpašības vārdu dzimti, skaitli un locījumu nosaka pēc lietvārda, ar kuru kopā tas lietots (*rozā kreklam* – īpašības vārds vīriešu dzimtes vienskaitļa datīvā).

2.4. Skaitļa vārds

Lokāmo skaitļa vārdu pamatforma ir vīriešu dzimtes vsk. (*viens, pirmais*) vai dsk. (*pieci*) nominatīvs.

Nelokāmajiem skaitļa vārdiem nepiemīt šādas gramatiskās kategorijas: dzimte un locījums (4. un 6. pozīcijas vērtības – 0).

2.5. Vietniekvārds

Vietniekvārda pamatforma parasti ir vsk. nominatīvs (*es, tu, nekāds, katra*). Vīriešu dzimtes vietniekvārdiem tas ir vīriešu dzimtē (*viņš, tāds, neviens*), sieviešu dzimtes – sieviešu dzimtē (*viņa, tāda, neviena*). Ja vietniekvārdam nav vsk., tā pamatforma ir dsk. nominatīvā (*abi*). Vietniekvārda sevis pamatforma ir vsk. ģenitīvs (*sevis*).

Noliegtie vietniekvārdi (pamatforma – *neviens, nekas, nekāds*) tiek aprakstīti kā nenoteiktie vietniekvārdi ar noliegumu (2. pozīcijas vērtība – i, 7. pozīcijas vērtība – y).

2.6. Apstākļa vārds

Apstākļa vārda pamatforma ir apstākļa vārda vienīgā forma (*blakus, aizgūtnēm*) vai pamata pakāpe (*labi*), ja apstākļa vārds tekstā lietots pārākajā (*labāk*) vai vispārākajā pakāpē (*vislabāk*).

Ja adverbam piemīt prievārda funkcija, resp., tas piesaista citus vārdus kādā locījumā (*blakus galdam, cauri mežiem, pāri Daugavai*), tie marķēti kā relatīvie adverbi (2. pozīcijas vērtība – r).

Versijas

Versija	Moduļa versija	Publiskās relīzes versija	Apraksts	Datums	Autors
0.1	5.5	UDv2.5	Pirmais apraksts – defisavienojumu dēļ.	2019-05-21	Laura
0.2		UDv2.5	Precizēts **spacing** lietojums	2019-10-18	Lauma
0.3		UDv2.9	Precizēta terminoloģija un pievienotas piezīmes par morfoloģisko marķēšanu	2021-11-15	Laura
0.4		UDv2.15	Sīkumi	2024-10-31	Lauma
1.0		UDv2.16	Pāriets uz atvērto dokumentu formātu	2025-04-28	Lauma