*Lauma Pretkalniņa, Laura Rituma
Mākslīgā intelekta laboratorija,
LU MII*

**Latviešu valodas sintaktiski marķētā korpusa tokenizācijas un morfoloģiskās marķēšanas metodikas apraksts**

# Tokenizācija

Kaut kāds ievads par pamatprincipiem, kā notiek tokenizācija – pamatā automātiski ar rīka palīdzību, bet dažos gadījumos mēs iejaucamies rezultātā, apvienojot vairākus teksta primitīvus vienā vai sadalot vienu teksta primitīvu divos. Kurā teksta slānī šī iejaukšanās parādās (vēl bez w līmeņa)? Kurā līmenī parādās iejaukšanās , ja ievieto vai izņem kādu teksta primitīvu kokā?

## Teksta primitīvi, kurus veido vairāki teksta elementi?

Kā viens teksta primitīvs 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.;
* Decimālskaitļi, kuriem pa vidu ir komats, piem., 14,57;
* Pulksteņlaiki, kuros izmanto punktu vai kolu, piem., 15:00, 15.00, 15:34:02.

Kā viens teksta primitīvs 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**.

Viens teksta primitīvs ir arī tādi saīsinājumi, starp kuru daļām iespējama atstarpe:

* skaitļi un telefona numuri (xn un xx), 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.

## Teksta elementi, kas sadalāmi vairākos atsevišķos teksta primitīvos

Dažos gadījumos automātiskā tokenizācija attēlo kā vienu teksta primitīvu tādu teksta elementu, kuru tomēr nepieciešams manuāli sadalīt vairākos teksta primitīvos:

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

## Tokenizācijas labojumu atveide morfoloģijas slānī (.m failos)

Tokenizācija atšķiras šādos gadījumos:

1. Ja oriģināltekstā ir atsevišķu burtu kļūdas,
2. Ja vārds oriģināltekstā kļūdas dēļ ir izrakstīts 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 kāds teksta primitīvs** (vārds vai pieturzīme), ir vēl citi lauki — tām ir viens **m** lauku bloks (apraksta morfoloģiju) un patvaļīgs skaits **w** bloku (katrs bloks apraksta vienu teksta primitīvu — piemēram vārdam „**3,14**” var būt trīs **w** bloki).

* **m/w/token** — tekstā lietotā vārdforma
* **m/w/no\_space\_after** — 1, ja starp doto teksta primitīvu 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 vairākiem teksta primitīviem (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 (vajag likt visas, kas atbilst):
	+ **spell** — izlabota drukas kļūda;
	+ **punct** — izlabota pieturzīmes kļūda;
	+ **insert** — ielikts jauns teksta primitīvs, kas tekstā bija izlaists;
	+ **spacing** — (a) aiz šī vārda oriģināltekstā trūkst vajadzīga atstarpe, (b) vairākos ar atstarpi izrakstītos tokenos izrakstīts vārds apvienots vienā vārdā;
	+ **union** — vārds sastāv no diviem vai vairākiem teksta primitīviem;
	+ **num\_normalization** — mainīts skaitļu pieraksts, piemēram, no AM/PM uz Latvijā lietoto, utt. — *vai šāds vispār datos ir lietots?*

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** pie pirmās daļām;
	+ kopā rakstāms vārds (“**jādara**”) uzrakstīts atsevišķi (“**jā dara”**) — **spacing** un **union**;
	+ nepareizi satokenizēts, 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 iekšā ir nepareiza atstarpe (**1 00** nevis **100**) ir nomarķēta kā **spell** — pareizrakstības kļūda.

# Morfoloģija

Šobrīd morfoloģiskās marķēšanas un lemmu izvēles principi ir iespiesti tagseta dokumentā SemTi-*Kamols\_morphotags.xlsx*

# Versijas

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Versija** | **Moduļaversija** | **Publiskās relīzes versija** | **Apraksts** | **Datums** | **Autors** |
| 0.1 | 5.5 | UDv2.5 | Pirmais apraksts — defissavienojumu dēļ. | 2019–05–21 | Laura |
| 0.2 |  | UDv2.5 | Precizēts **spacing** lietojums | 2019–10–18 | Lauma |