Korpuslingvistika

Kristel Uiboaed, Eleri Aedmaa, Maarja-Liisa Pilvik

2016 sügis

Korralduslikku. Kursuse materjalid

- Kursuse koduleht: `korpuslingvistika.ut.ee`

Korralduslikku. Kursuse materjalid

- Kursuse koduleht: `korpuslingvistika.ut.ee`
- Kursuse info ja materjalid kodulehel (ja Dropboxis)

Korralduslikku. Kursuse materjalid

- Kursuse koduleht: `korpuslingvistika.ut.ee`
- Kursuse info ja materjalid kodulehel (ja Dropboxis)
- Täielik kirjanduse loetelu on Zoteros.

Korralduslikku. Hinde kujunemine

- Kursuse hinne kujuneb

Korralduslikku. Hinde kujunemine

- Kursuse hinne kujuneb
  - jooksvalt tehtavate ülesannete põhjal

Korralduslikku. Hinde kujunemine

- Kursuse hinne kujuneb
  - jooksvalt tehtavate ülesannete põhjal
  - kursuse projekti lõpphinde põhjal (täpsemad juhised kursuse teises pooles)

Korralduslikku. Hinde kujunemine

- Kursuse hinne kujuneb
  - jooksvalt tehtavate ülesannete põhjal
  - kursuse projekti lõpphinde põhjal (täpsemad juhised kursuse teises pooles)
- Kõik kodused ülesanded ja projekt peavad olema esitatud

Korralduslikku. Hinde kujunemine

- Kursuse hinne kujuneb
  - jooksvalt tehtavate ülesannete põhjal
  - kursuse projekti lõpphinde põhjal (täpsemad juhised kursuse teises pooles)
- Kõik kodused ülesanded ja projekt peavad olema esitatud
- Tähtajast hiljem esitatud ülesanded punkte ei anna, kuid peavad olema esitatud ja positiivselt arvestatud

Korralduslikku. Hinde kujunemine

- Kursuse hinne kujuneb
  - jooksvalt tehtavate ülesannete põhjal
  - kursuse projekti lõpphinde põhjal (täpsemad juhised kursuse teises pooles)
- Kõik kodused ülesanded ja projekt peavad olema esitatud
- Tähtajast hiljem esitatud ülesanded punkte ei anna, kuid peavad olema esitatud ja positiivselt arvestatud
  - tagasisidet saavad ainult õigeaegselt esitatud ülesanded

Koralduslikku. Hinde kujunemine

- Kursuse hinne kujuneb
  - jooksvalt tehtavate ülesannete põhjal
  - kursuse projekti lõpphinde põhjal (täpsemad juhised kursuse teises pooles)
- Kõik kodused ülesanded ja projekt peavad olema esitatud
- Tähtajast hiljem esitatud ülesanded punkte ei anna, kuid peavad olema esitatud ja positiivselt arvestatud
  - tagasisidet saavad ainult õigeaegselt esitatud ülesanded
  - hiljem esitatud ülesandeid kontrollitakse kursuse lõpus ja võidakse saata parandamiseks tagasi (sellega tuleb ajaplaneerimisel arvestada)

Korralduslikku. Hinde kujunemine

- Kursuse hinne kujuneb
  - jooksvalt tehtavate ülesannete põhjal
  - kursuse projekti lõpphinde põhjal (täpsemad juhised kursuse teises pooles)
- Kõik kodused ülesanded ja projekt peavad olema esitatud
- Tähtajast hiljem esitatud ülesanded punkte ei anna, kuid peavad olema esitatud ja positiivselt arvestatud
  - tagasisidet saavad ainult õigeaegselt esitatud ülesanded
  - hiljem esitatud ülesandeid kontrollitakse kursuse lõpus ja võidakse saata parandamiseks tagasi (sellega tuleb ajaplaneerimisel arvestada)
- Loengutes ja praktikumides kohalviibimine pole kohustuslik, kuid kodused ülesanded eeldavad praktikumides ja loengutes omandatud oskusi ja teadmisi.

Korralduslikku. Koduste ülesannete esitamine

- Koduste ülesannete esitamine Dropboxi kaudu

Korralduslikku. Koduste ülesannete esitamine

- Koduste ülesannete esitamine Dropboxi kaudu
- Ülesannete esitamise tähtaeg on reedeti kell 17.00

Korralduslikku. Koduste ülesannete esitamine

- Koduste ülesannete esitamine Dropboxi kaudu
- Ülesannete esitamise tähtaeg on reedeti kell 17.00
- Koduste ülesannete faili nimetamine: perekonnanimi_ülesandeNumber_märksõna, näiteks: Uiboaed_2_sonadevaheliseSeose (palume mitte kasutada täpitähti)

Hindeskaala

A: 91 - . . .
B: 81-90
C: 71-80
D: 61-70
E: 51- 60
F: 0-50

Õppejõudude kontaktid (Jakobi 2-430)

- Kristel Uiboaed (kristel.uiboaed@ut.ee)
- Eleri Aedmaa (eleri.aedmaa@ut.ee)
- Maarja-Liisa Pilvik (maarja-liisa.pilvik@ut.ee)
- Konsultatsiooniaeg on reedeti kell 12:00 ruumis Jakobi 2-430

Lisaks

- Tony McEnery e-kursus keskkonnas Future Learn `https://www.futurelearn.com/courses/corpus-linguistics-2014-q3`
- Ja soovitav kirjandus

Mis on mis?

- **keeleteadus, lingvistika** - teadus keelest, selle olemusest, ehitusest, talitlemisest ja arenemisest
- **informaatika** - arvutil põhineva infotöötlusega tegelev teaduse ja tehnika haru

Mis on mis?

- **Arvutilingvistika, raallingvistika** (*computational linguistics*) loomuliku keele automaattöötlusega tegelev keeleteaduse ja informaatika piiriala
- **Keeletehnoloogia** (*language technology, natural language processing (NLP)*) tegeleb meetodite, tarkvara ja seadmetega, mis on spetsialiseeritud tekstide ja kõne töötlemiseks.
- KT on AL tehnoloogiline haru, mis tugineb teadmistele inimkeelest.
- Praktikas kasutatakse sünonüümidena.

Keeletehnoloogia koostisosad

- tarkvara: arvutiprogrammid keeleandmete töötlemiseks, nt
  - teksti grammatiline analüüs ja süntees
  - suulise kõne süntees ja tuvastus
  - õigekeelsuse ja stiili kontroll
  - masintõlge
  - infootsisüsteemid
  - dokumenditöötlus
  - inimkeelne dialoog arvutiga
  - tõlkija või keeleõppija abivahendid jne jne
- keeleressursid: formaalsed keeleandmed tarkvarasüsteemide arendamiseks:
  - elektroonilised sõnastikud ja andmebaasid
  - lingvistiliselt märgendatud tekstikorpused
  - formaalsed grammatikakirjeldused

Mis on korpus?

Keeleteaduses on sõna ***korpus*** all enne arvutite kasutuselevõttu tavaliselt mõeldud keeleainese kogumikku, mida kasutatakse uurimistöös materjalina (esineb see siis kartoteegi, lindikogu vms kujul) vastandina autori enda intuitsioonil põhinevatele üldistustele.

Mis on korpus?

Arvutiajastul: **polüfunktsionaalne elektroonilisel kujul olev tekstikogu**, millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi kogu keelest (selle hetkeseisust või muutumisest).

Aga tegelikult mida sageli korpuseks nimetatakse?

- Tekstikogu

Aga tegelikult mida sageli korpuseks nimetatakse?

- Tekstikogu
- Representatiivne tekstikogu

Aga tegelikult mida sageli korpuseks nimetatakse?

- Tekstikogu
- Representatiivne tekstikogu
- Representatiivne ja struktureeritud tekstikogu

Aga tegelikult mida sageli korpuseks nimetatakse?

- Tekstikogu
- Representatiivne tekstikogu
- Representatiivne ja struktureeritud tekstikogu
- Representatiivne, struktureeritud ja märgendatud tekstikogu

Korpuse representatiivsus

- Representatiivne – “esinduslik”
- Statistikas: representatiivne valim = populatsioonile tunnuste poolest vastav valim

- Üks eesti keele korpus peaks siis esindama

Korpuse representatiivsus

- Representatiivne – “esinduslik”
- Statistikas: representatiivne valim = populatsioonile tunnuste poolest vastav valim

- Üks eesti keele korpus peaks siis esindama
  - . . . kogu eesti keelt?

Korpuse representatiivsus

- Representatiivne – “esinduslik”
- Statistikas: representatiivne valim = populatsioonile tunnuste poolest vastav valim

- Üks eesti keele korpus peaks siis esindama
  - . . . kogu eesti keelt?
  - . . . mingit keelekasutust/allkeelt?

Allkeeled

- (2) Määruse alusel § 3 lõikes 1 nimetatud taotleja kaudu põllumajandustoodete töötlejatele ning taotleja liikmetele, kes ei ole põllumajandustootjad ega põllumajandustoodete töötlejad, antav toetus on vähese tähtsusega abi komisjoni määruse nr 1998/2006, milles käsitletakse asutamislepingu artiklite 87 ja 88 kohaldamist vähese tähtsusega abi suhtes (ELT L 379, 28.12.2006, lk 5–10), mõistes;

Allkeeled

- (2) Määruse alusel § 3 lõikes 1 nimetatud taotleja kaudu põllumajandustoodete töötlejatele ning taotleja liikmetele, kes ei ole põllumajandustootjad ega põllumajandustoodete töötlejad, antav toetus on vähese tähtsusega abi komisjoni määruse nr 1998/2006, milles käsitletakse asutamislepingu artiklite 87 ja 88 kohaldamist vähese tähtsusega abi suhtes (ELT L 379, 28.12.2006, lk 5–10), mõistes;

- icici sry kui ma veica kahtlane

Kuids saavutada representatiivsust?

Korpuse planeerimise etapil määrata kindlaks:

- korpusesse kuuluvad allkeeled
- ilmumisaja piirid
- selgitada välja (valitud allkeele) tekstiklasside osakaal (hulk ja mõju sellel perioodil)

Representatiivsuse näide: Brown ja LOB

TEKSTIKLASS	BROWN	LOB
Ajakirjandustekstid: reportaažid	44	44
Ajakirjandustekstid: juhtkirjad	27	27
Ajakirjandustekstid: ülevaate- ja probleemartiklid	17	17
Religioosne kirjandus	17	17
Harrastused, oskused	36	38
Populaarkirjandus	48	44
Biograafiad, esseed	75	77
Ametlikud dokumendid	30	30
Teaduslik kirjandus	80	80

Representatiivsuse näide: Brown ja LOB (järg)

TEKSTIKLASS	BROWN	LOB
"Üldine" ilukirjandus (*general fiction*)	29	29
Detektiivi- ja põnevuskirjandus	24	24
Ulmekirjandus	6	6
Seikluskirjandus, vesternid	29	29
Armastuslood	29	29
Huumor	9	9

Korpuste liigitusvõimalusi

- Suletud ehk representatiivne *vs* avatud ehk monitorkorpus
- Katkendikorpus *vs* tekstikorpus
- Kirjalik *vs* suuline keelekasutus (+ praegu ka multimodaalne korpus)
- Ükskeelne – kakskeelne – mitmekeelne
- Diakrooniline – sünkrooniline
- Üldkeel *vs* erialakeel
- Puhas tekst *vs* märgendatud

Korpuste põhjal tehtavate keeleteaduslike (uurimis)tööde liigitusvõimalusi (McEnery & Hardie 2012 järgi)

- Suhtlusviis, register
- Korpuspõhine *vs* korpusuuring
- Andmete kogumise viis
- Märgendatud *vs* märgendamata korpuse kasutamine
- Kõikehõlmav *vs* valikuline andmestik
- Mitmekeelne *vs* ükskeelne korpus

Mis on korpuslingvistika?

Termin kasutusel kahes tähenduses:

- **korpus**lingvistika

  Distsipliin, mis kirjeldab arvutikorpuste koostamispõhimõtteid ja rakendusvõimalusi.

Mis on korpuslingvistika?

Termin kasutusel kahes tähenduses:

- **korpus**lingvistika

  Distsipliin, mis kirjeldab arvutikorpuste koostamispõhimõtteid ja rakendusvõimalusi.

- korpus**lingvistika** ehk korpuspõhine lingvistika

  Lingvistika haru, mis paneb eriti rõhku tegeliku keelekasutuse uurimisele ja selle põhjal keelereeglite tuletamisele ning kvantitatiivsetele andmetele.

Milleks korpus?

Lingvistikas

- grammatikakirjeldused
- sõnavarauuringud
- keelestatistika
- sõna valents, sõnaühendid jne
- paralleelkorpused: võrdlev keeleuurimine, tõlketeooria
- keeleõpe: õppijate korpused
- keeleõpe: korpuse kasutamine võõrkeeleõppes

Kas iga tekstikogu on korpus?

- Tekstiarhiiv (nt ajalehel)
- Tekstoteek ehk elektrooniline raamatukogu (nt Eesti Kirjanduslugu Tekstides; Digiteeritud eesti ajalehed 1821-1944)
- Korpus – ühtses ja kontrollitud formaadis, metaandmetega varustatud, sageli lisainfoga varustatud (märgendatud)

Internet korpusena/korpuse asendajana (1)

Võimalused

Kasutada lihtsalt mõnd otsimootorit (*klikkama* vs *klikkima*)

- **Poolt**
  - lihtne ja kiire
  - Google tunneb nüüd ka eesti morfoloogiat (mitte küll täielikult, vesi : vee; susi : soe)
- **Vastu**
  - saab otsida ainult konkreetseid sõnu või fraase
  - Sagedusandmed kahtlased (mees *vs* mees – naine), ei näe tegelikult kõiki väidetavalt leitud esinemisi

Internet korpusena/korpuse asendajana (2)

Modifitseerida otsimootorisse suunatavat päringut ja vastuseks saadavat stringi

Google'i n-grammid

- Eesti keele jaoks oluline

Internet korpusena/korpuse asendajana (2)

Modifitseerida otsimootorisse suunatavat päringut ja vastuseks saadavat stringi

Google'i n-grammid

- Eesti keele jaoks oluline
  - keele tuvastamine (nüüd juba üsna hea)

Internet korpusena/korpuse asendajana (2)

Modifitseerida otsimootorisse suunatavat päringut ja vastuseks saadavat stringi

Google'i n-grammid

- Eesti keele jaoks oluline
  - keele tuvastamine (nüüd juba üsna hea)
  - täpitähed (ka juba parem)

Internet korpusena/korpuse asendajana (2)

Modifitseerida otsimootorisse suunatavat päringut ja vastuseks saadavat stringi

Google'i n-grammid

- Eesti keele jaoks oluline
  - keele tuvastamine (nüüd juba üsna hea)
  - täpitähed (ka juba parem)
  - morfoloogia

Internet korpusena/korpuse asendajana (2)

Modifitseerida otsimootorisse suunatavat päringut ja vastuseks saadavat stringi

Google'i n-grammid

- Eesti keele jaoks oluline
  - keele tuvastamine (nüüd juba üsna hea)
  - täpitähed (ka juba parem)
  - morfoloogia
  - terviklause või mõtestatud tekst

Hiljutisest diskussioonist postiloendis *corpora*

“there are many features that are not present in web documents, and that are important for users (linguists, text researchers and language technologists):

- spoken language
- dialects
- speech situations
- dialogue
- source and translated texts
- free choice of text types and genres
- grammatical annotation (and other linguistic annotation)
- background information on the text producers (age, gender, mother tongue, place of birth, place of living, education etc.)"

Internet korpusena/korpuse asendajana (3)

Veel võimalusi

- koostada korpus Internetis leiduvast materjalist
- koostada korpus automaatselt veebilehekülgi alla laadides ja korpuse kujule teisendades
- koostada korpus käsitsi või pool-käsitsi valitud lehekülgedelt kogutud tekste automaatselt korpuse kujule teisendades
  - Selleks arendatud spetsiaalseid tööriistu, nt WebBootCaT (`http://www.sketchengine.co.uk/`)

Leksikograafias (1)

- Korpusenäited sõnastikuartikli alusena
- Korpusenäited (lihtsustatud) näitelausetena sõnaraamatus
- Sõna erinevate tähenduste järjestamine sageduse järgi
- Uudissõnade tuvastamine monitorkorpuse abil

Leksikograafias (2)

- Sõnastiku koostamine (poolautomaatselt) paralleelkorpuse põhjal vt ka `http://kde.teataja.ee/`
- vt lisaks Tsepelina & Veskis 2010
- Võrdle verbide *minema* ja *tulema* wordsketch'e
- Kas leiad mingeid vigu? Millised võiks olla nende vigade esinemise põhjused?

Natuke korpuslingvistika ajalugu

- Eellugu: strukturaallingvistika (keeleteaduse valitsev suund 20. sajandi algupoolel), eriti selle suuna Ameerika haru (nim ka antropoloogiline lingvistika): uurisid seni kirjeldamata keeli nii, et koostasid selle keele korpuse ja selle põhjal tegid grammatikakirjelduse
- Generatiivne grammatika
  - Chomsky: *any natural language corpus is skewed*
  - Pidurdas korpuslingvistika arengut ligi 30 aastaks

Natuke ajalugu

Vastuväited Chomskyle

1. (Korpusest) kogutud keelematerjal on kõigile uurimiseks kättesaadav. Introspektiivsel vaatlusel põhinevaid järeldusi on palju raskem tõestada.
2. Introspektiivsed andmed on kunstlikud. Laused, mida kirjeldab introspektiivset meetodit kasutav lingvist, erinevad suurel määral lausetest, mis tüüpiliselt esinevad korpuses.
3. Inimestel on tavaliselt mingi sõna või konstruktsiooni esinemissagedusest ainult ähmane ettekujutus. Korpused on parimad seda liiki teabe allikad.

Natuke ajalugu

- Korpuste 1. põlvkond (*ca* 1960ndate lõpp – 1980ndate lõpp)

Natuke ajalugu

- Korpuste 1. põlvkond (*ca* 1960ndate lõpp – 1980ndate lõpp)
  - Brown ja Lancaster–Oslo-Bergen + palju analooge

Natuke ajalugu

- Korpuste 1. põlvkond (*ca* 1960ndate lõpp – 1980ndate lõpp)
  - Brown ja Lancaster–Oslo-Bergen + palju analooge
  - 1960ndate ameerika ja briti inglise keel nt Frown (90ndad samal printsiibil)

Natuke ajalugu

- Korpuste 1. põlvkond (*ca* 1960ndate lõpp – 1980ndate lõpp)
  - Brown ja Lancaster–Oslo-Bergen + palju analooge
  - 1960ndate ameerika ja briti inglise keel nt Frown (90ndad samal printsiibil)
- Korpuste 2. põlvkond (1990ndate algusest alates)

Natuke ajalugu

- Korpuste 1. põlvkond (*ca* 1960ndate lõpp – 1980ndate lõpp)
  - Brown ja Lancaster–Oslo-Bergen + palju analooge
  - 1960ndate ameerika ja briti inglise keel nt Frown (90ndad samal printsiibil)
- Korpuste 2. põlvkond (1990ndate algusest alates)
  - avatud monitorkorpused – koostaja vastutus $\rightarrow$ kasutaja vastutus

Natuke ajalugu

- Korpuste 1. põlvkond (*ca* 1960ndate lõpp – 1980ndate lõpp)
  - Brown ja Lancaster–Oslo-Bergen + palju analooge
  - 1960ndate ameerika ja briti inglise keel nt Frown (90ndad samal printsiibil)
- Korpuste 2. põlvkond (1990ndate algusest alates)
  - avatud monitorkorpused – koostaja vastutus $\rightarrow$ kasutaja vastutus
- Erandeid:

Natuke ajalugu

- Korpuste 1. põlvkond (*ca* 1960ndate lõpp – 1980ndate lõpp)
  - Brown ja Lancaster–Oslo-Bergen + palju analooge
  - 1960ndate ameerika ja briti inglise keel nt Frown (90ndad samal printsiibil)
- Korpuste 2. põlvkond (1990ndate algusest alates)
  - avatud monitorkorpused – koostaja vastutus $\rightarrow$ kasutaja vastutus
- Erandeid:
  - BNC (British National Corpus)

Natuke ajalugu

- Korpuste 1. põlvkond (*ca* 1960ndate lõpp – 1980ndate lõpp)
  - Brown ja Lancaster–Oslo-Bergen + palju analooge
  - 1960ndate ameerika ja briti inglise keel nt Frown (90ndad samal printsiibil)
- Korpuste 2. põlvkond (1990ndate algusest alates)
  - avatud monitorkorpused – koostaja vastutus $\rightarrow$ kasutaja vastutus
- Erandeid:
  - BNC (British National Corpus)
  - ANC (American National Corpus). Suured, uued ja suletud

Otsingud eri korpustest

- Enne korpuse kasutamist **ALATI, ALATI, ALATI**
  - Tutvu korpuse tutvustusega (mida see korpus sisaldab, milliseid tekste, mis ajastust, millist lisainformatsiooni on esitatud jne).
  - Tutvu korpuse kasutajaliidesega, selle juhendiga, otsinguvõimalustega.

Valimi koostamine korpuse materjali põhjal

- Väga oluline on läbi mõelda oma valimi koostamise kriteeriumid ja need enne materjali kogumist kindlalt määratleda (ja oma töös selgelt lahti kirjutada).

Valimi koostamine korpuse materjali põhjal

- Väga oluline on läbi mõelda oma valimi koostamise kriteeriumid ja need enne materjali kogumist kindlalt määratleda (ja oma töös selgelt lahti kirjutada).
- Kriteeriumid tuleb määrata eelkõige uurimisküsimust silmas pidades, kuid paratamatult tuleb arvestada ka olemasolevate võimalustega (korpused, otsinguvõimalused jmt).

Valimi koostamine korpuse materjali põhjal

- Väga oluline on läbi mõelda oma valimi koostamise kriteeriumid ja need enne materjali kogumist kindlalt määratleda (ja oma töös selgelt lahti kirjutada).
- Kriteeriumid tuleb määrata eelkõige uurimisküsimust silmas pidades, kuid paratamatult tuleb arvestada ka olemasolevate võimalustega (korpused, otsinguvõimalused jmt).
- Sageli valitakse x arv esimeste lauset, mis pole parim lahendus, kui korpuse otsingu tulemused väljastatakse allikate kaupa esinemisjärjekorras.

Sageduste normaliseerimine

- Paratamatus on, et sageli on meil kasutada piiratud materjalihulk, rääkimata esinduslikkusest ja tasakaalust (vt ka ptk 1.4.5).

Sageduste normaliseerimine

- Paratamatus on, et sageli on meil kasutada piiratud materjalihulk, rääkimata esinduslikkusest ja tasakaalust (vt ka ptk 1.4.5).
- Korpuslingvistilises uurimuses pole harv juhtum, kus meil tuleb võrrelda seda, mis pole objektiivselt võrreldav.

Sageduste normaliseerimine

- Paratamatus on, et sageli on meil kasutada piiratud materjalihulk, rääkimata esinduslikkusest ja tasakaalust (vt ka ptk 1.4.5).
- Korpuslingvistilises uurimuses pole harv juhtum, kus meil tuleb võrrelda seda, mis pole objektiivselt võrreldav.
- Näiteks pole otseselt võrreldav valim, mis on saadud 1000 ja 10 000 sõna suurusest korpusest. On ootuspärane, et 10 000 sõna hulgas esineb ka uuritavat vormi, konstruktsiooni jne rohkem kui 1000 sõna hulgas.

Sageduste normaliseerimine

- Paratamatus on, et sageli on meil kasutada piiratud materjalihulk, rääkimata esinduslikkusest ja tasakaalust (vt ka ptk 1.4.5).
- Korpuslingvistilises uurimuses pole harv juhtum, kus meil tuleb võrrelda seda, mis pole objektiivselt võrreldav.
- Näiteks pole otseselt võrreldav valim, mis on saadud 1000 ja 10 000 sõna suurusest korpusest. On ootuspärane, et 10 000 sõna hulgas esineb ka uuritavat vormi, konstruktsiooni jne rohkem kui 1000 sõna hulgas.
- Üks levinud võte korpuslingvistilistes töödes selle probleemi lahendamiseks on sageduste n-ö normaliseerimine, nt sagedus 100, 1000 jne sõna kohta. Normaliseerimisbaas võib olla ka miski muu (nt korpuse keskmine suurus, korpusfaili suurus vmt).

Sageduste normaliseerimine. Kuidas?

$$NormSagedus = (absoluutsagedus / korpuseSuurus) * normaliseerimisbaas.$$

- Näide
  - Oletame, et meil on kaks korpust suurusega 250 000 ja 400 000 sõna, mille põhjal soovime uuritavat nähtust võrrelda.
  - Uuritavat konstruktsiooni esines esimeses korpuses 324 korda ja teises 565 korda.
  - Kuidas need sagedused võrreldavaks saada?
    - (324/250 000)*100 000 = 130
    - (565/400 000)*100 000 = 141

Paralleelkorpus

- Korpus, mis sisaldab sama teksti vähemalt kahes keeles, mille laused, osalaused jne on paralleelistatud.
- Paralleelistamine on paralleelteksti üksteise tõlkeks olevate osade märgendamine, st näidatakse ära, millised laused, osalaused, fraasid, sõnad (kasutatakse harva) on omavahel vastavuses.
- Vabavaralised keelest sõltumatud paralleelistajad:
  - Vanilla aligner
  - Hunalign

Paralleelkorpuse näide

Lindgreni “Hulkur Rasmus” eesti-inglise

Ent samas eksitas ta mõtteid tasane hüüatus altpoolt:

But a hiss from below interrupted his deliberations.

Rasmus,

“Rasmus!

peitu!

Hide!

OPUSe eesti keelt sisaldavad allkorpused

- ECB European Central Bank corpus
- EMEA European Medicines Agency document
- EUconst The European constitution
- KDE4 KDE4 localization files
- KDEdoc the KDE manual corpus
- OpenSubtitles

Märgendamine

- Märgendamiseks nimetatakse interpretatiivse info lisamist suulist või kirjalikku keelt esindavasse keelekorpusesse.
  - Sisu: mida märgendada
  - Vorm: kuidas märgendada

Märgendamise näide: lause (1)

- Eesmärk: tunda ära ja märkida lausepiirid tekstis
  - Sisu: Mis on lause?
  - Mida selles korpuses / selle automaatse lausete märgendajaga käsitleda lausena?

Märgendamise näide: lause (2)

Mis on lause? (EKK):

- Lause on keelelise suhtluse põhiüksus.
- Tüüpiline lause sisaldab finiitset (pöördelist) verbivormi.
- Tüüpilised laused on lihtlaused, st laused, mis ei sisalda teisi lauseid. (*Sügis on käes. Vihma sajab.*)
- Lause, mille koosseisus on üks või mitu osalauset, on liitlause. (*Sügis on käes ja vihma sajab. Kass, keda ma nägin plangul kõõlumas, varitses varblasi, kes pahaaimamatult saiatükki nokkisid.*)

Märgendamise näide: lause (3)

EKK: Lause **kui struktuuriline tervik** ei pruugi kirjas alati kokku langeda **ortograafilise lausega, mis algab suurtähega ja lõpeb lauselõpumärgiga.**

Mida siis lausena märgendada? Mis on vajalik ja mis on võimalik?

Kirjalik tekst – ortograafiline lause (ja osalause)

Aga pealkirjad, vahepealkirjad, autorikirjed, tabelid, loendid jms?

Teksti segmenteerimine: lause (1)

Nt lõik ühest seadusetekstist

Vald ja linn võivad võtta laenu või emiteerida võlakohustust tõendavaid väärtpabereid järgmistel tingimustel:

1) kõigi laenude kogusumma ei või ületada 75% eelarvetuludest;

2) jne jne 3 lk;

. . .

x) laen võetakse valla või linna arengukavas ettenähtud investeeringuteks.

Teksti segmenteerimine: lause (2)

AMRITSARI KULDTEMPEL, SIKHI UNIVERSUMI KESKPUNKT

ÜLO SUURSAAR

Uhked värvilistes turbanites mehed, mõõgad vööl, sõjakas tuluke silmades helkimas. Selle stereotüübi aluseks on tõenäoliselt sikhid, India üks omanäolisemaid inimrühmi. Heade kavatsustega algus Sikhismi sünnimaa Punjab (Pandzab) on Induse keskjooksu . . .

Teksti segmenteerimine: laused (3)

Suuline keel:

J: kuidas nad on suutnud need nii õhukesed kõik saavutada (..) sajakolmekümne leheküljelised tavaliselt nad on ju

A: ja millega te tahate minna (..) kas te tahate seal (.) hh ee mingi transpordiga või või jalgrattaga või

Jututoad:

ma tean ma ise joonistasin selle

ma ei saa kindlat töökohta vastu võtta ma liiga andekas

Märgendamise näide: lause (4)

Vorm: kuidas esitada lausepiire tekstis?

Sügis on käes.@ Vihma sajab.@

Sügis on käes.

Vihma sajab.

Sügis on käes. Vihma sajab.

Märgendamise tasemed

1. Tehniline (või ka ortograafiline)
2. Morfoloogiline
3. Süntaktiline
4. Semantiline
5. Pragmaatiline

Tehniline märgendamine

Teksti struktuur:

nt ajalehes artiklid, nende pealkirjad, autorid, lõigud, laused, tabelid, väljajäetud materjal (nt tabelid, joonised, valemid jne)

Sõnad:

sõnapiirid, “mitmesõnalised sõnad”, nt *New York*, vihm *läks üle*, ma *ei tee* jpt

Morfoloogiline märgendamine (1)

Morfoloogia ehk vormiõpetus on grammatika osa, mis tegeleb sõnavormidega – nende moodustamise ja nendest arusaamisega.

Sõnavormide moodustamist käsitlevat vormiõpetuse osa nimetatakse vormimoodustuseks ehk morfoloogiliseks sünteesiks ja sõnavormidest arusaamist käsitlevat osa morfoloogiliseks analüüsiks

Analüüs: **raamatu/te/le**, **külla**

Morfoloogiline märgendamine (2)

Annab infot

- lemma ehk algvorm -> leksikonid
- sõnaliik
- morfoloogilised kategooriad (st käändsõnal arv, kääne; võrdlussõnal võrdlusastmed; tegusõnal pööre, tegumood, aeg, kõneviis ja kõneliik)

Slaidimaterjal toetub Kadri Muiscneki loengumaterjalidele.