Teadusteksti kriitiline lugemine & hinnang

DTI6001.DT Õppimine kõrgkoolis

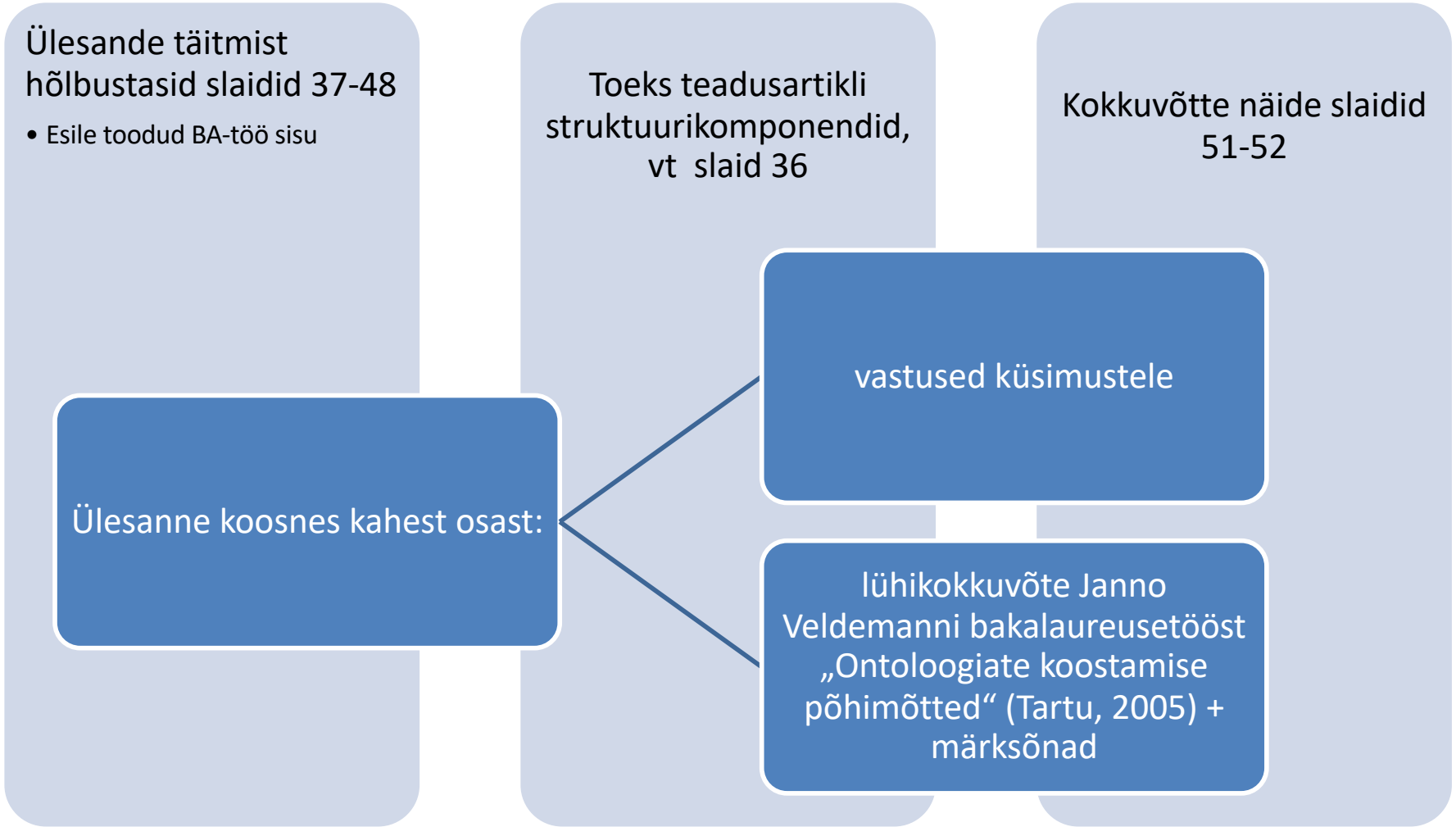
12.10.2018

Pille Eslon

Kava

- Tagasiside: ülesanne 1
- Ontoloogiate & keeletehnoloogia keskne roll kommunikatsiooniprotsessis
- Teadlik lugemine
  - Teema & valdkond
  - Sisulised märksõnad
  - Teadusteksti struktuur, ülesehituse loogika & keel
  - Analoogia & ülekanded

Tagasiside ja kordamine: ülesanne 1



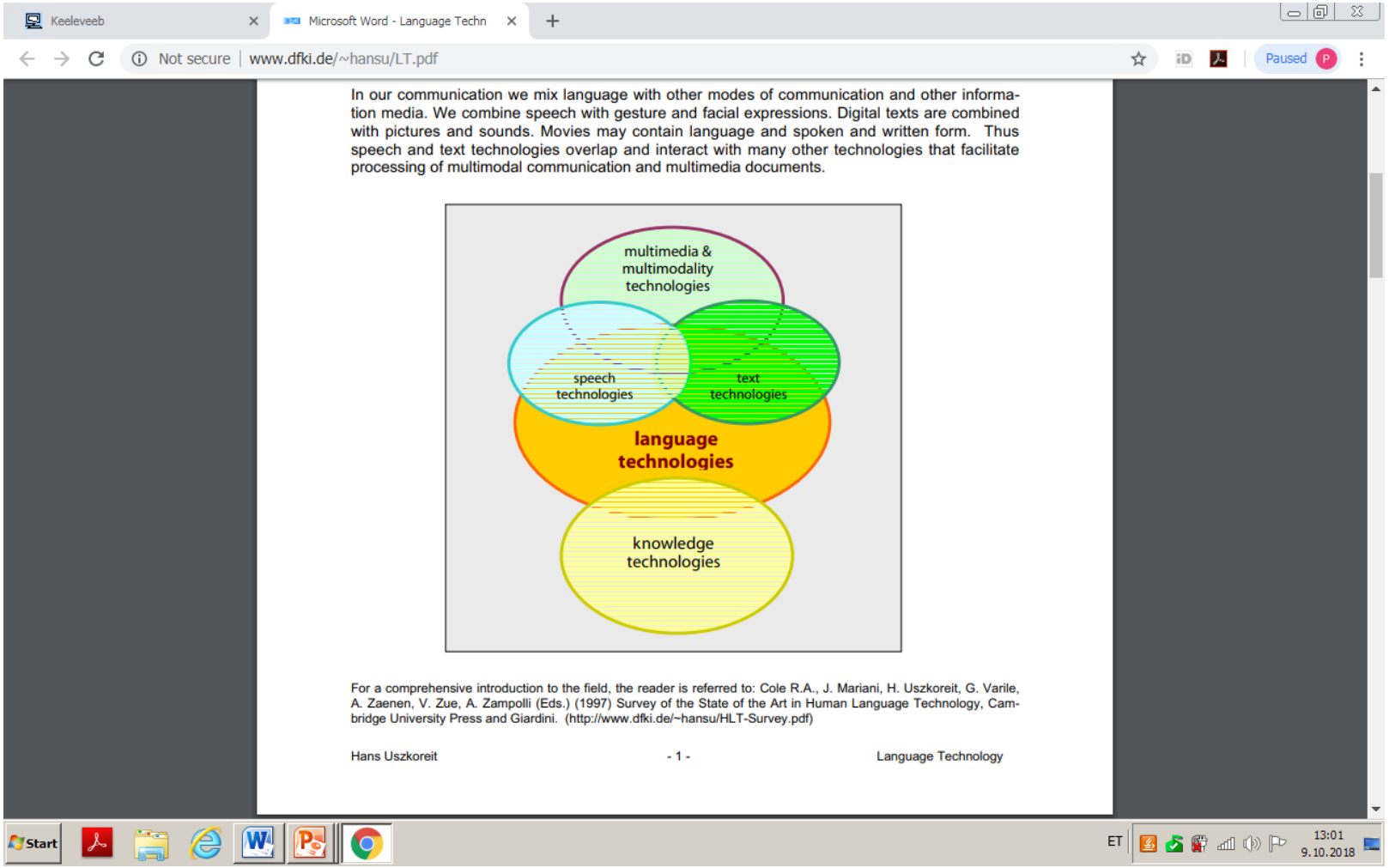
Näide

Kokkuvõte. Janno Veldemanni bakalaureusetöös „Ontoloogiate koostamise põhimõtted“ (Tartu, 2005) on antud ülevaade üldistest ja valdkondlikest ontoloogiatest, nende vajalikkusest, neile omastest tunnustest, loomise sammudest ja metodoloogiast (teoreetiline lähtepunkt). Autor on metodoloogia all silmas pidanud ontoloogia väljatöötamiseks mõeldud *juhtnööre* (*protsessijuhendit*) ning tuntumate seast valinud METHONTOLOGY. Sobivad arenduskeskkonnad on Veldemanni arvates Protégé, DAG-Edit/OBO-Edit ja Chimaera, sobivad esituskeeled aga Knowledge Interchange Format (KIF) ja Web Ontology Language (OWL).

Märksõnad: ontoloogia, metodoloogia, arenduskeskkond, esituskeel

Ontoloogiate ja keeletehnoloogia keskne roll kommunikatsiooniprotsessis

http://www.dfki.de/~hansu/LT.pdf



MASINAD

- OSUTAVAD TEENUST
  - aitavad suhelda (nt e-post, e-valimised, e-pank)
- TALLETAVAD TEADMUST & AJALOOLIST MÄLU
  - integreerivad ja vahetavad valdkondlikku teadmust
  - koguvad ja süstematiseerivad andmeid (erinevat liiki andmekogud, nt keelekorpused, e-sõnastikud, entsüklopeediad)
  - aitavad digiteerida ja säilitavad andmekogusid (pildi- ja filmiarhiivid, arhitektuuripärand, haruldased väljaanded jm)
- KODEERIVAD/DEKODEERIVAD INFOT
  - muudavad kirjaliku teksti (tähemärgid) suuliseks kõneks (heli) ja vastupidi
  - aitavad tõlkida loomuliku keele tekste
- TAOTLEVAD KASUTAJASÕBRALIKKE TEHNOLOOGIAID & RAKENDAVAD STANDARDEID
  - Loomuliku keele töötlemise vahendid
  - Tekstitöötlusvahendid

Loomuliku keele töötlemise vahendid

- Keelemudelid: statistilised, reeglipõhised ja segamudelid
  - Tarkvara arendamise standardid > VISL = Visual Interactive Syntax Learning (vt https://visl.sdu.dk/remoting.html)
    - POS-tagging (sõnaliigi märgendus)
    - Parsing (andmete liigendamine, nt sõltuvuspuud, vorm ja funktsioon)
- Keelemudelite rakendamine tekstide töötlemisel, et leida vajalikku infot
  - keelelist: semantika, morfoloogia, süntaks; keelekasutusmustrid
  - sisulist: nt teksti hinnangulisus, tonaalsus
    - nt „keele alkeemia“https://alchemy-language-demo.mybluemix.net/?cm_mc_uid=84202806432014811944632&cm_mc_sid_50200000=1481194463
    - nt emotsioonidetektor http://peeter.eki.ee:5000/applications/list

Tekstitöötlusvahendid

- Teksti kodeerimise (annoteerimise) standardid
  > TEI = Text Encoding Initiative
  (http://www.tei-c.org/)
- Tekstitöötlusvahendid kui ressurss
  - https://keeleressursid.ee/et/keeleressursid/tekstitootlusvahendid
  - Eestikeelse teksti sisukokkuvõtja
    http://lepo.it.da.ut.ee/~kaili/Syntax/estsum.html

Teksti lugemine

- Teadlik lugemine: valdkond & teema & motiivid
  - Motiiv: mis on ajendanud, ärgitanud teemaga tegelema?
    - objektiivsed asjaolud – vajan taustainfot
      - et kaardistada teemaga seotud eelnevad uurimused ja olemasolevad uurimissuunad
      - et leida uudne, aktuaalne ja originaalne vaatenurk uurimisprobleemi lahendamiseks jm
    - subjektiivsed asjaolud, nt isiklik huvi millegi/kellegi vastu, individuaalne seotus uurimisobjektiga

Näide

„Töö autor on puutunud kokku meditsiini valdkonna andmete kogumise tarkvara väljatöötamisega AS-is EGeen. Selle käigus on tekkinud vajadus paremini aru saada erinevate valdkondade võimalikest vajadustest. Praegu võimaldab AS EGeen välja arendatud tarkvaras kasutada suuri hierarhilisi mõistete süsteeme nagu ICD-9 ja ICD-10. Edaspidises arenduses aga vaja hakata arvesse võtma keerukamaid graafi-kujulisi ontoloogiaid ja muid meetode, kuidas integreerida mõistete süsteeme kasutajasõbralikku tarkvarakeskkonda, mida kasutab arvutikauge tavakasutaja, näiteks arst.“

(Veldemann 2005: 3)

- Hetkevajadus – eriala baasteadmiste omandamine
  - sh kirjalikud tööd, esitlused, esinemised, iseseisev uurimus- ja/või arendusprojekt (BA-töö)
    - Tegevus 1 – teadusteksti lugemisharjumus, oskus leida kiiresti vajalik info
    - Tegevus 2 – üldarusaam akadeemilisest kirjutamisest

Kiire otsing

- Teema & märksõnad, autorid & valdkonnad
  - eriakirjandus – tunnustatud teadusajakirjad, raamatukogude andmebaasid, Googel jne > loeng 19. oktoobril
- Leitud tekstid
  - esmane tutvumine tekstiga
    - vaatan tutvustust
    - loen sisukorda
    - leian kokkuvõttest olulised märksõnad
    - tutvun retsensiooni sisuga, lehitsen lugejate arvamusi

Näide

- Allikas: Uszkoreit et al. 1997
- Link: http://www.coli.uni-saarland.de/publikationen/softcopies/Uszkoreit:1997:OFT.pdf (8.10.2018)
- Avaneb kollektiivne monograafia *Survey of the State of the Art in Human Language Technology*
  - toimetajad: Ron Cole jt
  - Cambridge University Press and Giardini, 1997

Monograafia ülesehitus ehk struktuur

1. Sisukord > missugused teemad, kes autorid
2. Kolm eessõna > mis fookuses

   k.a monograafia eesmärgi ja struktuuri tutvustus
3. Kas sisaldab infot minu teemaga seotud märksõna(de) või huvipakkuvate autori(te) kohta?

Vt monograafia lõpus

1) märksõnaindeksid

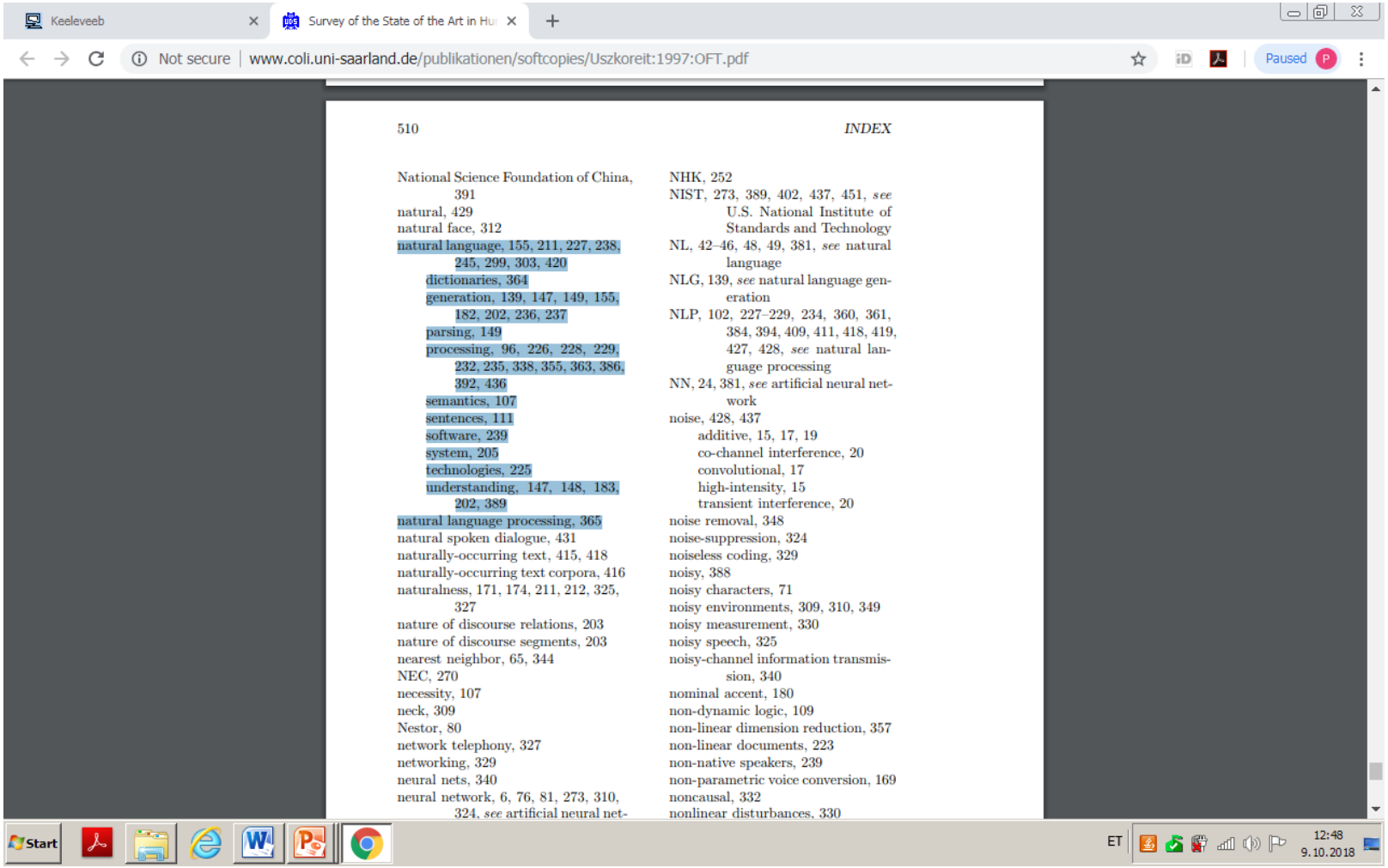
   nt *natural language processing* (loomuliku keele töötlus)

2) viidatud autorid

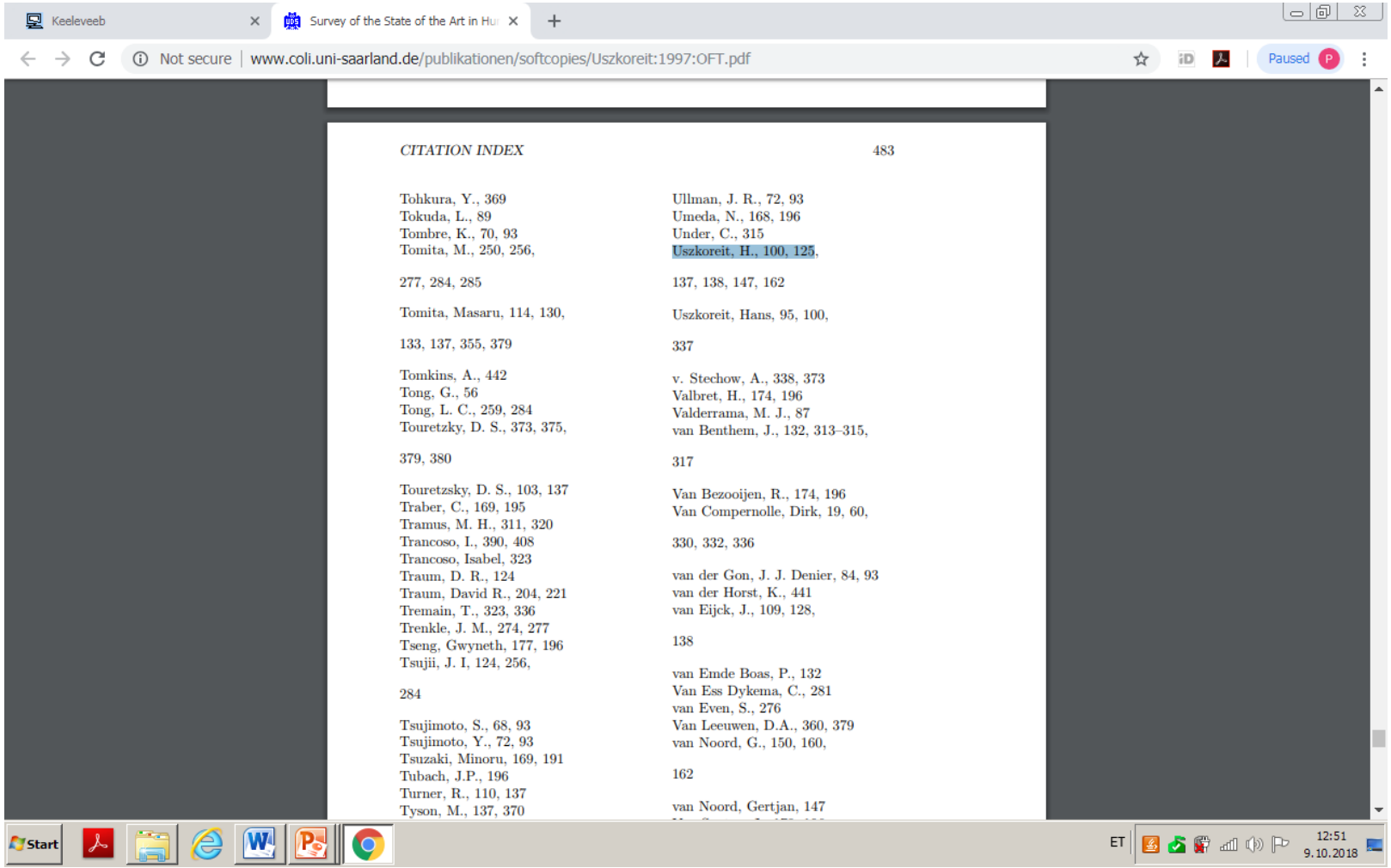
   nt Uszkoreit

Märksõna

natural language processing



Uszkoreit ~ Uzkoreit



Teadusteksti struktuur & ülesehituse loogika

- *Survey of the State of the Art in Human Language Technology* (1997) sisaldab
  - 13 peatükki, lühendite sõnastik, märksõnaindeks, viidatud autorite indeks
- Peatüki struktuur **keeleressurside näitel**
    - Sissejuhatav ülevaade
    - Kirjaliku keele korpused
    - Suulise keele korpused
    - Leksikonid
    - Terminoloogia
    - Keelekorpuste nimistu koos linkidega
    - Kirjandus

Peatükk keeleressursid: sissejuhatus

- Sissejuhatav ülevaade kaardistab „olukorra riigis“
  - **Antakse mõiste seletus**: „The term linguistic resources refers to (usually large) sets of language data and descriptions in machine readable form, to be used in building, improving, or evaluating natural language (NL) and speech algorithms or systems“ (Godfrey, Zampolli 1997: 381 jj).
  - **Näited**: „Examples of linguistic resources are written and spoken corpora, lexical databases, grammars, and terminologies, although the term may be extended to include basic software tools for the preparation, collection, management, or use of other resources“.

Peatüki sisu, teema aktuaalsus, vajadused

- **Peatüki sisu**: „This chapter deals mainly with corpora, lexicons, and terminologies“.
- **Teema aktuaalsus**: „An increasing awareness of the potential economic and social impact of natural language and speech systems has attracted attention, and some support, from national and international funding authorities. Their interest, naturally, is in technology and systems that work, that make economic sense, and that deal with real language uses (whether scientifically *interesting* or not)“.
- **Vajadused**: „This interest has been reinforced by the success promised in meeting such goals, by systems based on statistical modeling techniques such as hidden Markov models (HMM) and neural networks (NN), which learn by example, typically from very large data sets organized in terms of many variables with many possible values“.

Probleem(id)

- **Probleem 1**: „A key technical factor in the demand for lexicons and corpora, in fact, is the enormous appetites of these techniques for structured data. Both in speech and in natural language, the relatively common occurrence of relatively uncommon events (triphones, vocabulary items), and the disruptive effect of even minor unmodeled events (channel or microphone differences, new vocabulary items, etc.) means that, to provide enough examples for statistical methods to work,
- **Probleem 2:** the corpora must be numerous (at the very least one per domain or application), often massive, and consequently expensive“. „The high cost of creating linguistic resources.“

Rakenduslik tähendus, tulevikunõuded

- **Rakenduslik tähendus**: „... cooperative efforts of companies, research institutions and sponsors ... . This obviously requires that linguistic resources not be restricted to one specific system, but that they be reused—by many users (shareable or public resources), or for more than one purpose (multifunctional resources).“
- **Nõuded keeleressurssidele**:
  - „... linguistic resources must also be *theory-neutral*...“
  - „... a consensus among different theoretical perspectives and systems design approaches...“
  - „... the adoption of common specifications and *de facto* standards in creating linguistic resources and ensures their harmonization at the international and multilingual level“
    - » The Text Encoding Initiative (TEI) „... has produced a set of guidelines for encoding texts“

Tegevussuunad

- **Kolm tegevussuunda**:
  - converting existing linguistic resources to common standards; putting some resources in the public domain; to promote the reuse of existing linguistic resources
  - to promote the development of new linguistic resources for those languages and domains where they do not exist yet
  - to create cooperative infrastructure to collect, maintain, and disseminate linguistic resources on behalf of the research and development community
- **Protsessi käivitamine**: The most appropriate way to organize these activities is still under discussion in various countries

Lähitulevik

- **Valik konkreetseid tegevusi**:
  - collection of speech and text data
  - creation of Common Lexical Databases with free commercial licenses for members
  - publication of existing corpora previously available only to government contractors
  - intellectual property rights to existing linguistic resources
  - release of government-owned resources to researchers

The need for ensuring international cooperation in the creation and dissemination of linguistic resources seems to us a direct consequence of their infrastructural role, precompetitive nature, and multilingual dimension

Tänane seis: nt CLARIN

- CLARIN = Common Language Resources and Technology Infrastructure
  - the mission to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences
    - **digital language data** (in written, spoken, or multimodal form) for scholars in the **social sciences** and **humanities**
    - **advanced tools** to discover, explore, exploit, annotate, analyse or combine such data sets, wherever they are located
      - https://www.clarin.eu/content/services (CLARINi teenused)

Sissejuhatuse komponendid

- Teema tõstatub **objektiivsetest asjaoludest ja objekti olemusest**
- **Teoreetilis-rakenduslikud vajadused ja olulisus**
  - määravad uuringu **aktuaalsuse**
  - tõstatavad rea **küsimusi ja hüpoteese**
  - tingivad **eesmärgi** sõnastamise laiemas ja kitsamas kontekstis
  - suunavad objekti käsitlemise **vaatenurga ja meetodi(te)** valikut
    - teoreetiline lähtepunkt, töövahendid, esituskeel
  - avaldavad mõju töö **loogilisele ülesehitusele**

Näide Veldemann 2005 põhjal

- **Teema tõstatumine tingitud rakenduslikust vajadusest:** infot tuleb süsteemselt ja korrektselt hallata; selleks vaja saavutada ühiseid arusaamu
- **Võimalik lahendus**: võtta kasutusele selgelt defineeritud mõistete süsteemid e. ontoloogiad
- **Teema aktuaalsus**: organisatsioonide vaheline suhtlus vajab ühtset keelt, et möödarääkimiste osakaal oleks võimalikult väike; ontoloogiad organiseerivad ja võimaldavad ka tarkvarasüsteemide vahelist suhtlust
- **Teema jätkusuutlikkus (perspektiivsus)**: Ontoloogiatena kirja pandud informatsiooni saab taaskasutada, sellele saab ehitada nii organisatsioonisisese kui välise suhtluse

Järgneb

- **Eesmärk**: teha kindlaks, mis on ontoloogiad ja kuidas neid kasutada
  - **EBAKORREKTNE: eesmärgiks on uurida** *vs.* **eesmärk on kaardistada, seletada, põhjendada, kirjeldada** ...
- **Uurimisküsimused**:
  - Kus on võimalik ontoloogiaid kasutada?
  - Milliseid eeliseid annab nende kasutamine?
  - Millest ontoloogiad koosnevad, millist kuju võivad saada?
  - Mis on ontoloogite väljatöötamise eeldused ja vahendid?

Järgneb

- **Ajend ehk motivatsioon**
  - tuleneb praktilisest vajadusest arendada meditsiini valdkonna andmete kogumise tarkvara ja selle kasutuskeskkonda > rakendusvajaduste alla (vt eespool)
- **Töö ülesehitus ehk struktuur**
  - sisaldab iga peatüki sisu lühikirjeldust
  - tähelepanu sõnastamisele:
    - selgitatakse, kirjeldatakse, jaotatakse, vaadeldakse, antakse ülevaade, kirjeldatakse
      - **VÄLTIDA**: *peatükk vaatleb, annab ülevaate, kirjeldab* – ülekantud tähendus > publitsistlikum, kõnekeelsem
    - on selgitatud, on kirjeldatud, on antud ülevaade ...
    - *mina*-positsioonilt: selgitan, kirjeldan, annan ülevaate...

Ülesanne 3: Teadusartikli kriitiline hinnang

Lugege sissejuhatus läbi, hinnake vajalike sisukomponentide olemasolu ja otstarbekust slaidide 25-28 alusel. Kirjaliku hinnangu pikkus 1000 sõnakasutust ilma tühikuteta.