Verbmobili suulise keele puude pangad

Helen Nigol

3.05.2007

Mis on Verbmobil?

- Verbmobil oli masintõlke projekt, mille eesmärgiks oli välja töötada kõnelejast sõltumatu süsteem suulise kõne tõlkimiseks.
- Projekti oli kaasatud saksa, inglise ja jaapani keel.
- Kindlad teemad: reisi planeerimine, kokkusaamiste määramine ja kaug PC hooldamine.

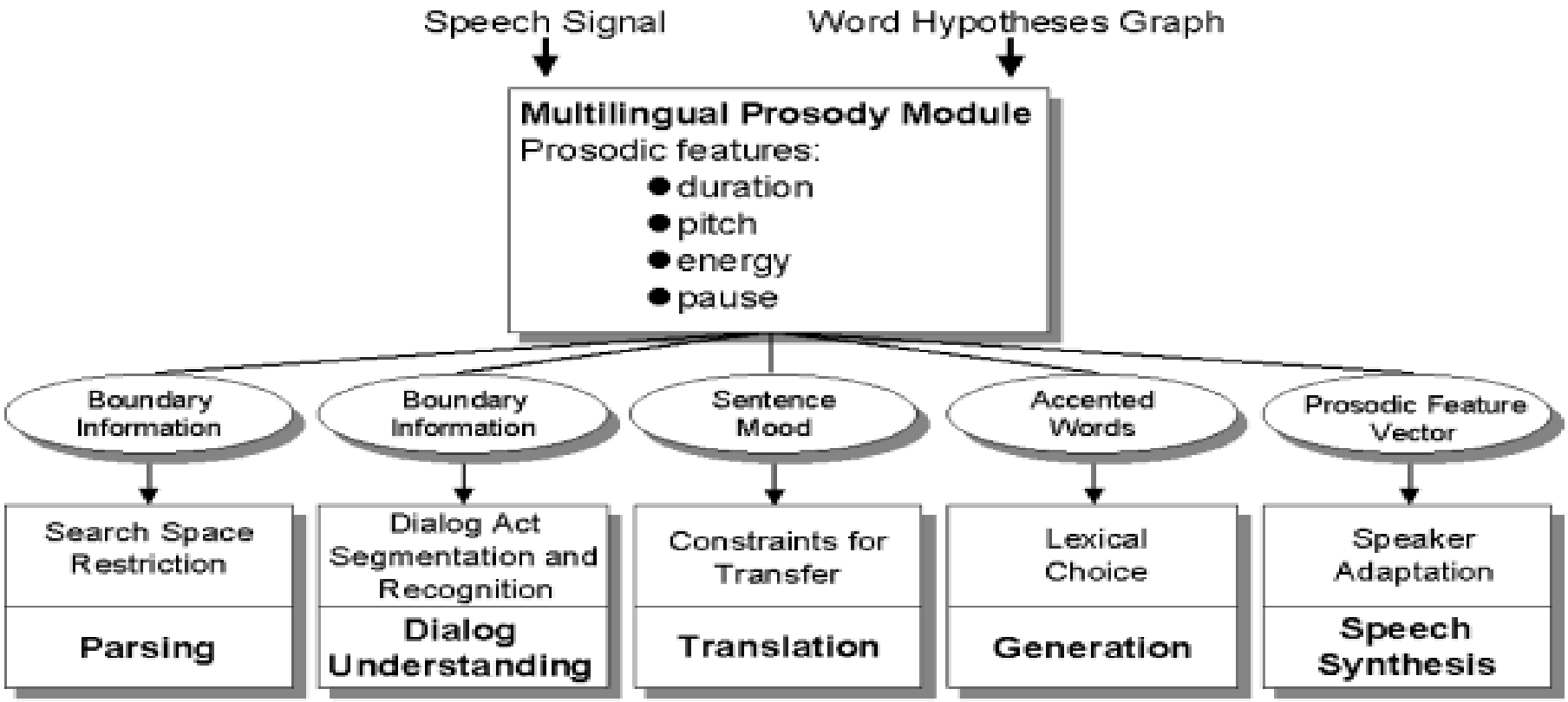
Mis erilist võrreldes teiste tõlkesüsteemidega?

- Vastupidiselt eelnevatele dialoogi tõlkesüsteemidele, mis tõlkisid teksti lausehaaval, on VS tõlge kontekstitundlik.
  - nii tõlgitakse saksakeelse sõna *nächste* inglise keelde vastavalt sellele, kas see väljendab lauses aega (“Millal läheb järgmine rong”) või kohta (“Kus asub lähim hotell”)
  - *Es geht bei mir. - „See sobib mulle küll“ või „Me võime kohtuda minu juures“*

Võtmesõnaks: prosoodia

- Prosoodilist informatsiooni kasutatakse süstemaatiliselt VS igal töötlustasandil.
- Prosoodia mooduli tulemusi kasutatakse parsimisel, dialoogi mõistmisel, tõlkimisel ja genereerimisel ja kõne sünteesil.
- Prosoodilised erinevused ühes keeles võivad vastata leksikaalsetele ja süntaktilistele teises keeles, nt saksakeelne kõne *wir haben noch...* tõlgitakse inglise keelde kas *we still* või *as we have another*, sõltuvalt sellest, kas *noch* on rõhutatud.

Võtmesõnaks: prosoodia (2)



Probleemid

- Kõne tuvastamisel tekkinud vead

Öeldud string: *when would be a good time for us to meet*

Tuvastatud string: *one would be a good time for us to meet*

- Tekkinud vead viivad selleni, et vaja pidevalt üle küsida ja informatsiooni täpsustada.
- Verbmobil on vahendaja, mistõttu ei saa ise küsimusi esitada.

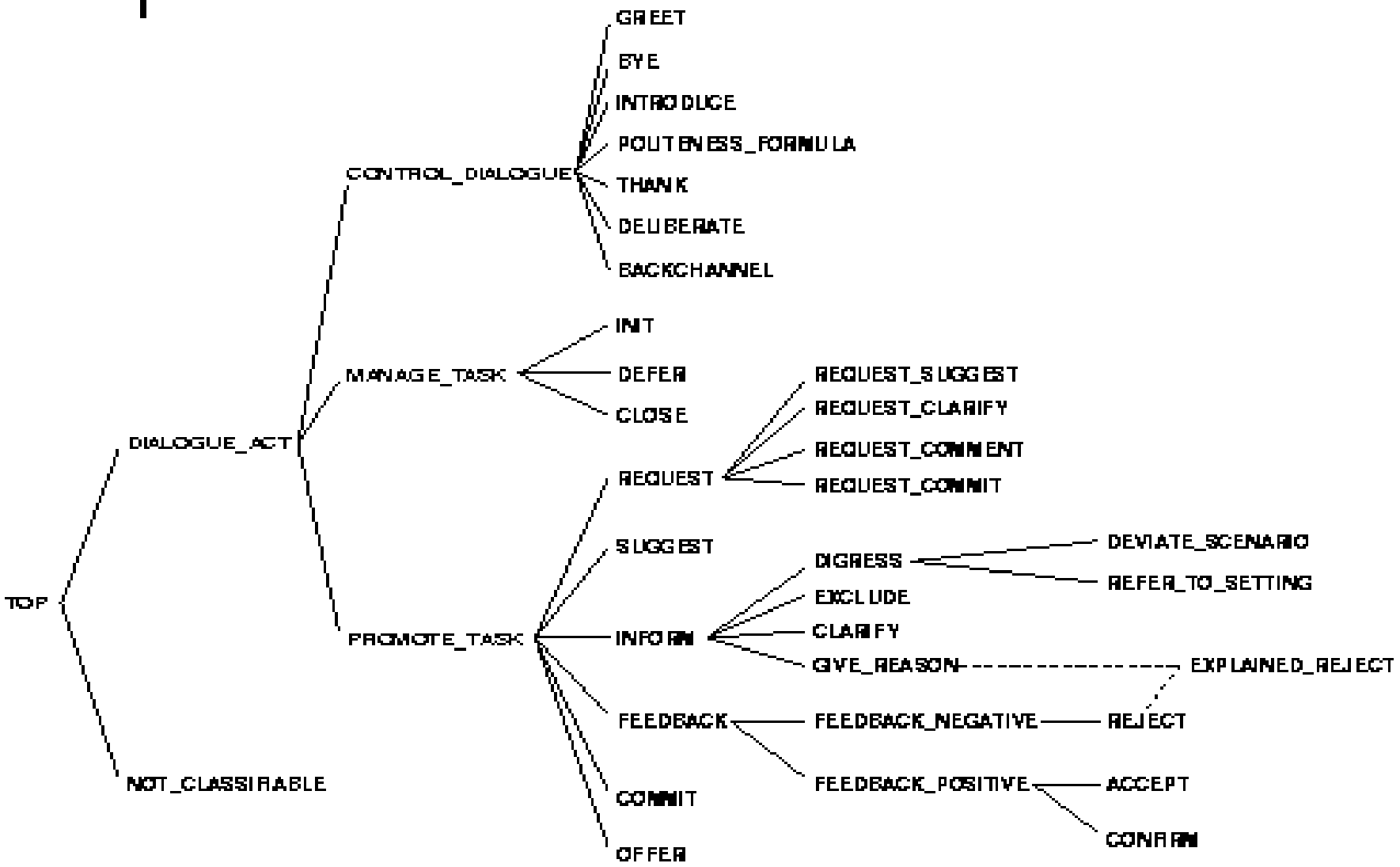
Verbmobil arvudes

- I faas 1993-1996 ja II faas 1997-2000
- I faas jaotati 16 allprojektiks ja 135 väiksemaks tööülesandeks, millega tegeles 33 uurimisrühma 28 ülikoolist, lisaks 125 inimest uurimiskeskustest ja erinevatest kompaniidest ühe aasta kohta
- II faasis jaotati töö 8 allprojektiks ja 127 tööülesandeks, mida lahendasid 24 erinevat töörühma
- Verbmobili kõne andmebaas sisaldab 3200 dialoogi: 1454 saksakeelset, 726 inglisekeelset ja 1020 jaapanikeelset. Erinevaid kõnelejaid oli kokku 1658: 1022 sakslast, 202 inglast ja 434 jaapanlast
- 7 aasta jooksul anti välja 238 Verbmobili raportit kogumahus 5331 lehekülge
- Verbmobili projekti rahastas *Bundesministerium für Bildung und Forschung*. Kokku on sellesse projekti raha paigutatud 116 miljoni saksa marga väärtuses.

Lauselised üksused

- Punkte ja komasid asendavad suulises kõnes mõne sõna rõhutamine ja fraasid ühest nn hingetõmbest teiseni.
- Sõnade jada *Ja-zur-Not-geht-es-auch-am-Samstag* (jah-vajaduse-korral-saab-ka-pühapäeval) on võimalik tänu erinevale rõhuasetusele analüüsida kaheti.
  - ”Ja, zur Not geht es auch am Samstag”
  - ”Ja, zur Not! Geht es auch am Samstag”

Kõneaktid Verbmobilis



Materjal

- Verbmobili kõnekorpust eristab teistest samalaadsetest see, et dialoogid on lindistatud kasutades mitut erinevat mikrofoni. Iga kõneleja hääl on salvestatud mitme erineval kaugusel asetseva mikrofoniga ja erinevate telefonidega. Seda kõike tehakse seepärast, et kõne äratundjat treenida materjaliga, millel on erinev audio signaali kvaliteet.
- Dialoogid Verbmobili kõnekorpuses on läbinud 15 erinevat märgendusetappi: kaks erinevat transliteratsiooni, leksikaalne ortograafia, kanooniline häältdus, käsitsi ja automaatne fonoloogiline segmenteerimine, sõna segmenteerimine, prosoodiline segmenteerimine, dialoogiaktid, müra, emotsionaalne kõne, süntaktilised kategooriad, sõna kategooriad, süntaktilised funktsioonid, prosoodilised piirid.

Verbmobili puude pangad

- Saksa keele puude pank (~38 000 puud)
- Inglise keele puude pank (~30 000 puud)
- Jaapani keele puude pank (~20 000 puud)

Märgendamine

lausetasand	peasõlme märgendid erinevat tüüpi lausetele
väljatasand	sõlme märgendid topoloogilistele väljadele
fraasitasand	sõlmemärgendid süntaktilistele kategooriatele ja servamärgendid grammatilistele funktsioonidele
leksikaalne tasand	sõnaliikide märgendamine

Märgendamine (2)

- Stylebook for the German Treebank in Verbmobil (118 lk)
- Stylebook for the English Treebank in Verbmobil (77 lk)
- Stylebook for the Japanese Treebank in Verbmobil (84 lk)

Märgendamine (3)

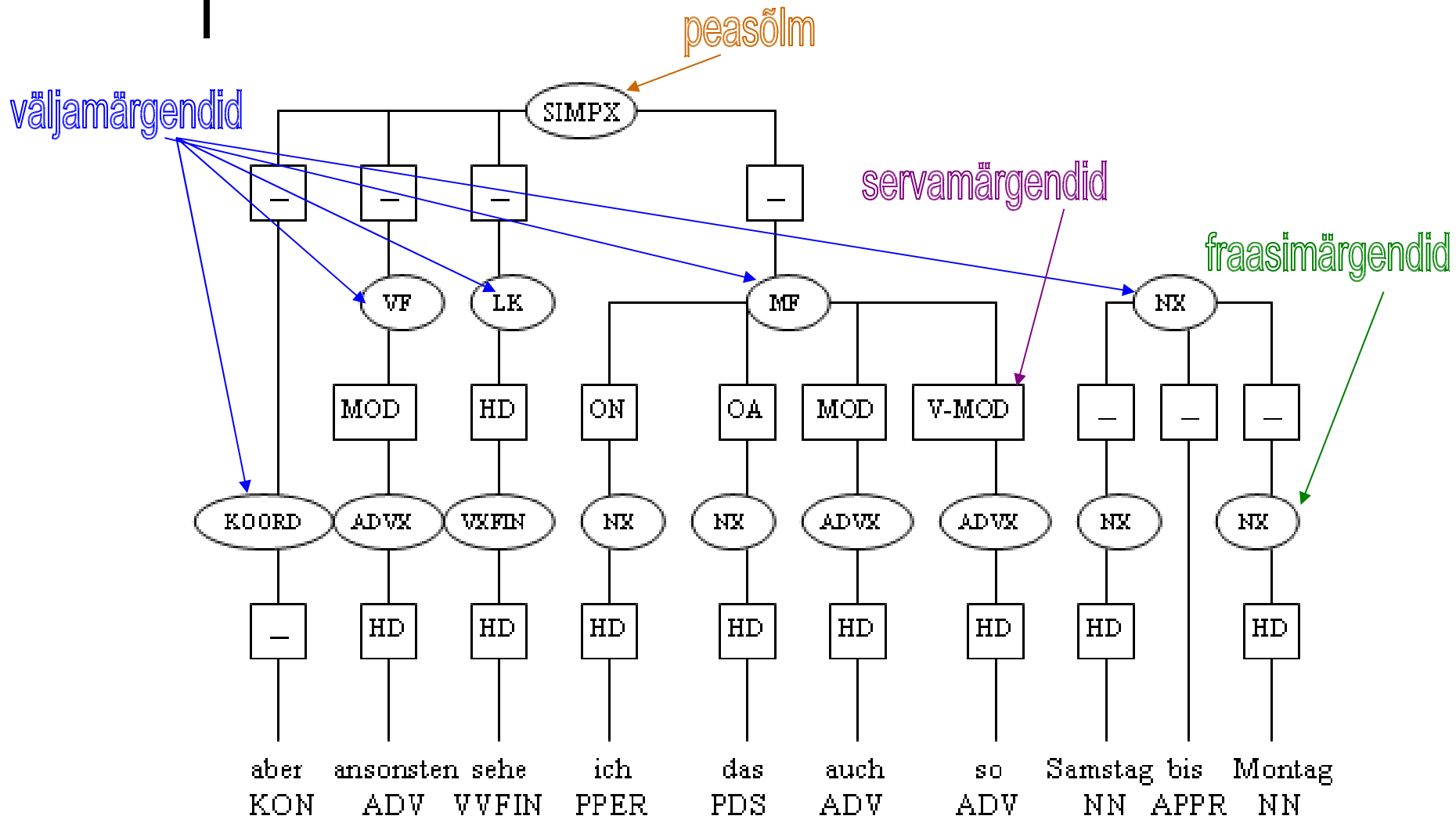
Underlying Design Principles

- Linguistically adequate annotation schemes
  - based on empirical syntax research
  - to ensure reusability of the data
- Structures that are easy to process
  - pure tree structures, no crossing branches
  - surface structure: no empty categories, no traces
  - flat constituent structures
- Consistency of annotation
  - semi-automatic annotation
  - automatic consistency checks

Printsiibid

- **Pikima ühtivuse printsiip** (ingl *longest match principle*) nõuab, et kõik tütarsõlmed oleksid seotud ühe vanemsõlmega nii, et analüüsitud moodustaja oleks nii süntaktiliselt kui ka semantiliselt hästi moodustatud.
- **Lameda klasterdamise printsiip** (ingl *flat clustering principle*) hoiab hierarhia tasandite numbrit süntaktilises struktuuris nii madalal kui võimalik.
- **Kõrge seostatavuse printsiip** (ingl *high attachment principle*) kirjutab ette, et mitmesed modifitseerijad seotakse kõrgeima võimaliku tasandiga süntaksipuus.

Süntaksipuu



Süntaksipuu (2)