CLARINO WP6 – Korpuskel-integrering

Paul Meurer

Uni Computing

Oslo, 4. juni 2012

Oversikt

1. Korpuskel – en nydesigned fleksibel korpusplatform
2. Clarino WP6: Integrasjon og videreutvikling
3. Feide-integrering

Outline

Hovedtrekk

- Modellert etter Corpus Workbench (CWB), men:
- Unicode (BMP = UCS-2)
- Støtte for hierarkiske data (XML)
- Innebygd støtte for flerverdiattributter (for å kode ambiguitet i grammatisk annotasjon og i metadata) og attributter med mengder som verdier (uordnete mengder: morfologiske trekk)
- Støtte for store korpus (2 milliarder posisjoner og mer)
- Rask evaluering av søk (raskere enn CWB), basert på statiske mmap-ete indeksfiler
- Nye algoritmer:
  - Tilstandsautomatene evalueres fra kanter med færrest antall korpusposisjoner
  - Regulære uttrykk på suffiksvektorer for raskt leksikonsøk

Korpuskel: Hovedtrekk

- Ekspressiv søkesyntaks, ligner på CQL (CWB), men bedre håndtering av flerverdi- og mengdeattributter og hierarkiske data
- Integrert søkbar korpusannotering og -editering, implementert som et ekstralag i en relasjonsdatabase
- Kobling til lyd og eksterne ordbøker
- Web-grensesnitt:
  - Konkordanser, kollokasjoner, distribusjonsstatistikk, ordlister, dokumentvisning, nedlasting mm.
  - Grafisk søkegrensesnitt
  - Web-rammeverk med wiki, lokalisering, administrering av brukere og rettigheter
- Skrevet i Common Lisp

Korpuskel: Prosjekter og korpus

Utvikling av kjernemodulen:

- finansiert av Meltzer-stiftelsen

Brukes for:

- ASK (Norsk andrespråkskorpus)
- Talebanken (dialektkorpus, med integrert kobling til lyd)
- Norsk aviskorpus
- Georgisk korpus (pilotversjon for Georgisk nasjonalkorus)
- mm.

Planlagt innenfor CLARINO:

- Utbygget Norsk aviskorpus
- COLT
- COLA
- ICAME (delvis)

Outline

Clarino WP6: Integrasjon

- Integrasjon i et federert autentiseringsrammeverk (Feide), i koordinasjon med INESS-prosjektet
- Metadata-portal
- Grensesnitt for selvregistrering og opplasting av nytt materiale
- Klargjøring av programvaren for distribusjon
- Interoperabilitet med LAP (WP9)

Clarino WP6: Videreutvikling

Utbygging og forbedring av funksjonalitet, bl.a.:

- Støtte for alle populære nettlesere (til nå: Chrome, Safari, Firefox, Opera; IE mangler)
- Parallellisering av noen søkealgoritmer for å utnytte regnekraften til HPC-klynger
- Effektivisering av indekseringsrutinene
- Bedre bruker- og API-dokumentasjon
- Funksjonalitet for parallelle korpus
- Bedre support for multimedia
- Kobling til statistikkpakken R

Outline

Feide - Felles Elektronisk IDEntitet

“Kunnskapsdepartementets valgte løsning for sikker identifisering i utdanningssektoren”

Feide forenkler prosessen ved å ta i bruk føderert identitetshåndtering:

- En bruker **registrerer** seg én gang: nemlig hos sin egen verts-organisasjon. Vertsorganisasjonen gir brukeren ett brukernavn og passord, og er ansvarlig for å vedlikeholde brukerens person-opplysninger. Vertsorganisasjonene er universiteter, høgskoler,...
- **Autentisering** gjøres alltid av vertsorganisasjonen, som også gir tjenestene eventuelle personopplysninger. Slik er alle Feide-tjenester tilgjengelige for brukerne med ett brukernavn og passord. Tjenestetilbyderne slipper å registrere nye brukere, fordi de får de opplysningene som trengs direkte fra brukernes vertsorganisasjoner.
- Avgjørelsen om en bruker skal få **tilgang** til tjenesten er basert på de opplysningene tjenesten får fra vertsorganisasjonen.

Feide-integrering

Feide-integrasjonen omfatter to deler:

- En administrativ del
- En teknisk del

Feide-integrering - administrativt

Administrativt:

- Ta kontakt med Feide og få registrert en testkonto som tjenestelverandør (support@feide.no)
- Utveksle metadata
- Teste tjenesten
- Finne ut behovene: hvilke attributter trenger tjenesten?
- Supplere tjenestebeskrivelse og logo til Feide
- Tegne kontrakt med Feide
- $\longrightarrow$ Tjenesten kan gå i produksjonsmodus

Feide-integrering - teknisk

Teknisk:

- Protokoll: SAML 2.0 (Security Assertion Markup Language) via https/ssl (XML-based)
- Installere SAML 2.0-kompatibel programvare: (Open source: SimpleSAMLphp, Shibboleth 2.0, Lasso, Mod_mellon osv.)
- Integrere i web-tjenesten
- Implementere protokollen mellom IdP (identity provider, = feide.no) and SP (service provider)
- Implementere håndteringen av to scenarioer (via HTTP redirect):
  - Login:
    - brukeren logger seg inn fra én SP og er autentisert for alle (federation)
    - Henting av attributter for autentiserte brukere, koble til lokal brukerdatabase og rettigheter
  - Single logout:
    - utgående LogoutRequest to SP, innkommende LogoutRequest from SP -> bruker kan velge å logge seg ut av én eller alle tjenester (SP)