Computational Linguistics for Low-Resource Languages

November 2, 2011

Alexis Palmer

UNIVERSITÄT DES SAARLANDES

Today

- ✦ scheduling, wiki, requirements, questions
- ✦ language resource assessments
- ✦ Abney & Bird 2010 (if time)

Nachrichten/News

- groups.google.com/group/cl4lrl -- email list (cl4lrl@googlegroups.com) and collaborative documents
- wiki.coli.uni-saarland.de/cl4lrl/main -- CoLi-hosted course wiki
- requirements -- 4/7CP options; questions?

Topics and scheduling

Topics and scheduling

November

- 09: NO MEETING!
- 16: Grammar engineering & Grammar Matrix
- 23: more on data
  - Human Language Project, 7 dimensions, IGT (me)
  - Data model for HLP, encoding wordlists (?)
  - GOLD (General ontology for lxc. description) (?)
- 30: morphology, rule-based
  - leveraging by mapping data (Ehsan?)
  - cross-linguistic adaptation of morphological analyzer: Xhosa/Zulu (Mariya?)

Topics and scheduling

December

- 07: morphology, unsupervised
  - Goldsmith, Morfessor (?)
  - newer approaches: alignment/projection (Iliana?)
- 14: POS tagging
  - POS tag induction (Peter?)
  - Universal POS tags (?)
- 21: syntactic parsing, projection/leveraging
  - Xia and Lewis, using IGT (?)
  - other cross-linguistic approaches (Jelke?)

Topics and scheduling

January/February

- 11: typological implications
  - inducing typological implications (Marc)
  - using implications for grammar induction (?)
- 18: language families
  - inducing familial relationships (Richard?)
  - using lg. phylogeny for grammar induction (?)
- 25: machine translation
  - crisis MT (i.e. rapid deployment) (?)
  - something else related to MT (?)
- Feb 1: other topics
  - cross-lingual IR (Birgit?)
  - TBD (?)

Topics and scheduling

November

- 09: NO MEETING!
- 16: Grammar engineering & Grammar Matrix
- 23: more on data
  - Human Language Project, 7 dimensions, IGT (me)
  - Data model for HLP, encoding wordlists (?)
  - GOLD (General ontology for lxc. description) (?)
- 30: morphology, rule-based
  - leveraging by mapping data (Ehsan?)
  - cross-linguistic adaptation of morphological analyzer: Xhosa/Zulu (Mariya?)

Topics and scheduling

December

- 07: morphology, unsupervised
  - Goldsmith, Morfessor (?)
  - newer approaches: alignment/projection (Iliana?)
- 14: POS tagging
  - POS tag induction (Peter?)
  - Universal POS tags (?)
- 21: syntactic parsing, projection/leveraging
  - Xia and Lewis, using IGT (?)
  - other cross-linguistic approaches (Jelke?)

Topics and scheduling

January/February

- 11: typological implications
  - inducing typological implications (Marc)
  - using implications for grammar induction (?)
- 18: language families
  - inducing familial relationships (Richard?)
  - using lg. phylogeny for grammar induction (?)
- 25: machine translation
  - crisis MT (i.e. rapid deployment) ()Philip
  - something else related to MT (?)
- Feb 1: other topics
  - cross-lingual IR (Birgit?)
  - TBD (?)

Language Resource Assessments

UNIVERSITÄT
DES
SAARLANDES

Languages

North America

- Cree (Mariona)
- Yurok (Richard)

Africa

- Xhosa or Ndebele (Mariya)

Asia

- Hokkaida Ainu (Antonia)
- Angami (Liling)
- Farsi (Ehsan)
- Kurdish (Ilyas) [+Europe]

Languages

Europe

- Tsakonian Greek (Nikos)
- Ladin (Iliana)
- Basque (Birgit)
- Irish (Andreas)
- Sorbian (Peter)
- Rhine Franconian, aka"Saarbrücken-Saarländisch" (Michael)
- Nordfriesisch (Philip)
- West Frisian (Jelke)
- German Sign Language (Marc)

German Sign Language 1

Data/linguistic resources/tools/other

- signed languages are not universal
- relationships have most to do with language teaching
- 80K Deaf speakers in Germany, 120K non-Deaf
- DGS is *not* just signed German
- uses classifiers (?) [give-paper vs. give-cup]
- 1880 claim made that DGS is *harmful* to Deaf Germans; 2002 finally designation of DGS as a foreign lg, allowing free access to translators
- large dialectal variation, esp. in domains of e.g. technical terminology, colors, country names, days of the week
-

German Sign Language 2

Data/linguistic resources/tools/other

- project building corpus of DGS/dialects (Hamburg)
- Hamnosis notation scheme, written sign
- some annotated resources, but not much
- Hamburg corpus will be linked to dictionary (or dictionary to corpus)
- wiki dictionary
- some computational projects
- privacy concerns (anonymity via avatars)

Cree (Eastern) 1

Data/linguistic resources/tools/other

- Cree is Algonquian language spoken in Canada, ~97K
- Eastern Cree ~12K, in Quebec and surroundings
- “macrolanguage”: dialect continuum wrt intelligibility
- was forbidden language for a long time
- currently: initiatives for rescuing the language
- current status: vulnerable but still being transmitted to younger generations
- primary data: translations of religious texts (3 Bibles, collections of songs and other religious texts)
- 2 alphabets: Roman alphabet, Cree syllabics (19th century)

Cree (Eastern) 2

Data/linguistic resources/tools/other

- Current movement to support use of syllabics
- Another domain with resources: education, but documents not available online
- There are some dictionaries, grammars, not easy to determine to which dialect given resources refer

Sorbian 1

Data/linguistic resources/tools/other

- Slavic language (same family as Czech & Polish)
- Eastern Germany, Western Poland
- estimated # of speakers: 18K Upper Sorbian, 7K Lower Sorbian
- Sorbian Institute in Kottbus & []
- institute hosts archive, bibliography
- several bilingual dictionaries exist, with German as reference language
- new dictionary in progress: ~60K keywords, meant to be used in schools
- also a phrase/idiom dictionary
- two searchable corpora

Sorbian 2

Data/linguistic resources/tools/other

- Lower Sorbian: News corpus, 23M tokens (!), 1848-1937
- Upper Sorbian: newer news (?) corpus
- both corpora are searchable
- there is a textbook online for self-teaching, also covers linguistics, history, culture
- 2nd source: U Leipzig, dictionary, ~100K sentences, this includes some ontological information
- Lexilogos: French web service, Declaration of Human Rights

Hokkaido Ainu

Data/linguistic resources/tools/other

- spoken in northern Japan (island of Hokkaido), formerly in some parts of Russia
- at present: 10 or fewer speakers (15 in 1996)
- traditional culture was essentially subsumed by dominant Japanese culture, with ethnic/cultural/linguistic differences ignored
- at some point Ainu were given some sort of protection as a culture and language
- there is some effort to revive the language
- one newspaper published in Ainu
- dictionary with sound files, some interlinear text
- reference language is generaly Japanese

Kurdish 1

Data/linguistic resources/tools/other

- 4th most commonly-used language in the Middle East
- ~10M speakers in Turkey, ~5M in the west, more in Iraq, Syria, Lebanon, Armenia, Iran [check]
- ~16M active speakers (Wikipedia)
- ethnic Kurd population ~25-30M people
- 2nd official language in Iraq, but not in other countries
- many dialects: 2 of these more dominant than others
- several different alphabets exist, Latin most common, also an alphabet similar to Arabic
- Kurdish Institute of Paris; Brussels; Stockholm; other cities

Kurdish 2

Data/linguistic resources/tools/other

- non-concatenative morphology, dual gender
- some linguists treat Kurdish as a dialect of Farsi, but this is controversial
- certainly closer to Persian than to Turkish
- quite a lot of material in Kurdish online
- not much in the way of NLP resources (i.e. corpora, etc.)
- there have been (or are still?) attempts to create a national corpus of the language
- Kurdish-Turkish, Kurdish-English, Kurdish-Farsi

Xhosa 1

Data/linguistic resources/tools/other

- not actually a minority language; one of 11 official languages of South Africa, also spoken in Lesotho
- ~8M speakers (1995 estimate) in South Africa
- tonal language, with click consonants
- highly agglutinative (typical of Bantu languages)
- language is quite vital: music, films, television, Wikipedia (though very limited), how-to-make-a-click-sound videos on YouTube
- currently has some sort of protected status: there is a governmental organization in South Africa supporting multilingualism (Pan-African Language Board)
- is taught at schools

Xhosa 2

Data/linguistic resources/tools/other

- not much data for Xhosa-English pair
- old grammar keeps being republished
- there are some parallel/comparable corpora with the other official languages, from S. African universities
- wordlists for spell-checking applications
- CALL resource for learning
- African WordNet exists, but it’s not so easy to access or find information about
- there is also some spoken corpus of telephone conversation (probably 10-20 hours/language)
- there is work leveraging lxc. closeness with other Bantu languages

West Frisian 1

Data/linguistic resources/tools/other

- spoken in the north of the Netherlands
- ~500K native speakers
- quite close to English, also to Dutch and Danish
- official language in its province (since 1954)
- 8 dialects, 4 major, 4 minor
- status: vulnerable
- regional TV station, radio stations, newspapers
- there is some West Frisian literature, also used in government and education
- Wikipedia (22K articles)
- archiving effort as well as a research institute

West Frisian 2

Data/linguistic resources/tools/other

- dictionary, with online version
- dictionaries and word lists for some of the larger dialects
- even as early as 1913: linguistic study of West Frisian
- some work on sociolinguistic aspects of lg. situation
- computational efforts re: West Frisian in their earliest stages
- diachronic corpus under development, more annotation for historical versions than for modern
- also there's a spoken corpus under development
- there's also some word-level linking to Dutch
- TTS by adapting system for Dutch (2004)

Angami 1

Data/linguistic resources/tools/other

- name means “devil’s tongue, devil’s language”
- spoken in NE India, in Nagaland
- 135K first-language speakers (Ethnologue, 2009); number increasing (?)
- what linguists call Angami is actually an artificial standardization for documentation
- conflicting documentation efforts: internal and external
- vulnerable status, though is taught in school, as a (second) language class, thus there are textbooks
- most resources from efforts of Christian missionaries

Angami 2

Data/linguistic resources/tools/other

- missionaries made a big deal about vocabulary to do with hunting (and therefore killing)
- language is written in a Latin alphabet
- grammars are conflicting, and w/out supporting texts
- there are decent lexicons, phrase books, etc.
- existing resources not amenable to machine reading
- there’s also a body of song lyrics in Angami
- songs last 3-10 days per performance
- exam papers, syllabi, PhD in Angami language
- audio resources are open but not free

Tsakonian Greek 1

Data/linguistic resources/tools/other

- dialect of Modern Greek, spoken by <300 people, mostly elderly, SE Pelopennesia [check]
- diverged from Modern Greek many centuries ago
- status: critically endangered (no longer being learned)
- some revival efforts from 1990s haven't persisted
- speakers not too interested in preserving the language
- interesting linguistic features have prompted study
- only written resources are transcriptions by linguists, these mostly are living in private libraries, etc.
- nothing really online or digitized
- there is a dictionary with some story transcriptions

Tsakonian Greek 2

Data/linguistic resources/tools/other

- there are some song transcriptions
- no systematic organization of resources
- one 3-minute recording found online
- supposedly there are some recordings at U Chicago
-

North Frisian (Nordfriesisch)

Data/linguistic resources/tools/other

- protected minority language in Germany
- dialect continuum of 10 dialects, with significant variation, also in written forms even for shared words
- virtually no computational resources, only native speakers, some dictionaries, schoolbooks
- only one dialect actively spoken (?)
- seriously endangered, but active dialects are quite active
- 2004 (?) legislation gave lg. some official status
- there is some literary tradition, both by native speakers and not (translations??)
- Bible (New Testament), Wikipedia (~1K)

Yurok

Data/linguistic resources/tools/other

- approximately 12 speakers, though there are ~5K members enrolled in the Yurok tribe
- located in California, Altic language (some relationship to Algonquian)
- Berkeley involved in some study, preservation
- currently: Master-Apprentice programs, projects in local schools, language classes, etc.
- Yurok Language Project (tribal program, University program): lots of educational materials and such
- active documentation efforts, both for language and cultural information (ethnographic information)
-