Less Ressourced Languages & BLARK

- The BLARK concept was defined in a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) [1]
- Defines a (minimal) set of LRs to be made available for as many languages as possible and mapping the actual gaps that should be filled in so as to meet the needs of the HLT field.

Resources founded for:

- Middle high German (1050 -1450)
- New high German (1500 -1700)
- Old English (500 – 1300)
- Middle new English and Renaissance
- Medieval Spanish (1200 -1400)
- Medieval French (1300 -1500)
- Latin
- Ancient Greek
„Monolingual“ Lexicons

- Most of the existent monolingual resources are in fact word-translators from the historical language into the modern one.
- Often associated with:
  - Morphological explanations
  - Listing of variants
- Not always it is transparent for the user which texts/paper-lexicons are the basis for these resources
http://drw-www.adw.uni-heidelberg.de/drw/

Seuche nach brückmeier i im Index fakesimiierte Brückentüten

Brückmeier I S. 5

marc da fe sich die geltigkeit bey der saf sich die
stuhlheiz wendem und so den man zu achten
Die man abe die den burger selbst heft die
ob gutter, da he seinen burger nicht akzept,
so fal die burger alle sol mit stuhlheiz leist als
sich gefühlt hat.* Abzog v dem leben der Erwähnt
(v. seinem) — „allmählich der was abbrunne
wurden von der christenheit“ — E. Abziehe,
Abziehen, Abzählen.
Abe und mit seiner Gesellschaft unter Ar.

http://wandership.ca/projects/
eow/

EOV: Englisches Onstligende Wordboc

Verb Inflections for *swendan

Swendan (v.), to change.

Infinitive: swendan
Inflected infinitives: to swendo
Present participles: *swendo
Past participles: *swended

<table>
<thead>
<tr>
<th>Person</th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>First person</td>
<td>Second person</td>
<td>Third person</td>
</tr>
<tr>
<td>Present</td>
<td>swendo</td>
<td>swendo</td>
</tr>
<tr>
<td>Past</td>
<td>swendo</td>
<td>swendo</td>
</tr>
<tr>
<td>Future</td>
<td>swendo</td>
<td>swendo</td>
</tr>
</tbody>
</table>

http://www.oldenglishtranslator.co.uk/

index.htm
Bi-lingual Lexicons

- Found were only:
  - Ancient Greek → English
  - Latin → English
  - English → Latin

- Most part of the Tools developed within the
  PERSEUS Project (http://www.perseus.tufts.edu/)

- For Ancient Greek and Latin also Morphological
  analysers are available
Conceptual Space

- Are difficult to realise because:
  - There is few information about the conceptual representation in middle age
  - Less specialisation
  - Differences between sacral and liturgical works

http://mhdbdb.sbg.ac.at:8000/

Middle High German
Syntactical structures

- Dependency Treebanks in development for Ancient Greek and Latin at the Perseus Digital Library Project
- Syntactic annotations in German diachronic corpora

Word distribution

- Renaissance English Collection (Perseus)
  - English (12,343,974 words)
  - Greek (1,183 words)
  - Latin (20,545 words)
  - Italian (6 words)
  - German (590 words)
  - French (5,365 words)
  - Other (24 words)

- Middle high English Collection (Perseus)
  - 452,162 words EN
  - 22,638 words

Corpora

- Perseus Digital Library (ancient Greek, Latin, early/middle English, Renaissance English, Old Norse)
- TLG (http://www.tlg.uci.edu/) (most literary texts written in Greek from Homer to the fall of Byzantium in AD 1453)
- Diachronic German Corpora (Humboldt University Berlin)

Dedicated Machine Translation Systems

- Latin – English
- „Brutus“: Latin to English Machine Translation - A Direct Approach (Pau Bowden – Nottingham University)
  - direct-MT attempt at Latin to English translation.
  - Incorporates a Latin morphological processor and two specialised English morphological processors,
  - produces translations on a sentence by sentence basis in a series of stages.
  - The first stage is a word for word transliteration, each English word being marked with case, gender, tense etc. information inherent in the Latin source word.
  - Subsequent stages correct English noun number and verb endings, find connected noun-adjective groups, mark subject and object, adverbial word order to standard English VSO
Any Chance with Google?

- Source Text (XIXth century English): It would seem fitting for a Northern folk, deriving the greater and better part of their speech, laws, and customs from a Northern root, that the North should be to them, if not a holy land, yet at least a place more to be regarded than any part of the world beside (Volsunga Saga. William Morris and Einir Magnusson, London. Walter Scott Press. 1888.)

- Google German Translation: „Es hat den Anschein, passend für eine Nord-Folk, im Zuge der stärkeren und besseren Teil ihrer Rede, die Gesetze und Gebrauche aus einer Nord-Wurzel, dass die Nord sollte zu ihnen, wenn nicht ein heiliges Land, noch mindestens einen Platz mehr zu gilt als Teil der Welt an."

Any Chance with SMT?

- Modern less resourced languages can be handled with:
  - Language Models trained on a large parallel corpora
- Such corpora are rare also for some modern languages.
- Recent experiments proved that the SMT quality decrease significantly when the domain of the test data differs from the training data
- If the training corpora is small the language model is not powerful enough
Distributed Resources

- The resources for historical languages presented before are distributed all over the world.
- For most part of them the processes behind the interfaces are not transparent to the user.
- Reusability through Web-Services

Other Languages?

- Many historical languages have no resources at all.
- Especially for these languages it is very important that the manuscripts become known to the European (world) scientific community.
- What about: church Slavonic, Arabic documents, northern languages.
- What about medieval versions of Latin and byzantine Greek?

What „BLARK” do we need to understand such texts?

http://www.romanichart.com/
books/texturas1.pptx.html

THE AGE OF ENRIQUE IV (1454-1474)