زمانەوانیی کۆمپیوتەری

بابەتی 2: سەرچاوە زمانەوانییەکان

سینا ئەحمەدی

sinaahmadi.github.io/KurdishCL

ناوەڕۆک

دەسپێک


NUDANSK
ORDBOG
HANDBOG
© Sina Ahmadi

سەرچاوە زمانەوانییەکان بۆ زمانەوانیی کۆمپیوتەری

- پێداوەی زمانەوانیی ← یەکەم هەنگاو بۆ پڕۆسەس کردنی زمان
- بەشێکی زۆری زمانەکانی دنیا سەرچاوەی زمان و زمانەوانییان بۆ کەمە [1]
- زمانی کوردی زمانێکی کەم سەرچاوەیە (less-resourced)

بەکارهێنانی سەرچاوەکانی زمان

- ناسینەوەی بەشەکانی وشە (part-of-speech tagging)
- ڕاست کردنەوەی هەڵەی ڕێنووس (spelling error correction)
- لێک دانەوەی واتای وشەکان: «**داریێکی** بە دەستەوە بوو.» (word-sense disambiguation)
- ناونانی ڕۆڵە ماناداڕەکان (semantic role labeling)
- پەیوەندی دانی چەمکەکان لە ڕستەدا (entity linking)
- وێڕای پێشکەوتنەکان لە بواری ژیریی دەستکرددا، سەرچاوەکانی زمان هێشتاش ڕۆڵێکی گرینگیان لە تەکنۆلۆژیای زماند[1]

سەرچاوەی زمانەوانی/زمانی

سەرچاوەی زمانەوانی بریتییە لە پێکهاتەی بەرهەمە زمانەوانی و زمانییەکان کە بۆ چێ کردن، باشتر کردن یان هەڵسەنگاندنی تەکنۆلۆژیای زمان و تۆژینەوەی سەر زمان بە کار دەهێنردرێن.

1. **پێدراوەی زمانەوانی**: ئەو سەچاوەیانەی زمانێک لە ڕێگای پێدراوەوە دەناسێنن، وەک فەرهەنگ، لیستی وشە یان تەنانەت دەنگی شێوەزارێک
2. ئەو **ئامێرانەی** کە بۆ بەکارهێنان، نیشان دان یان دەست تێ وەردانی پێدراوەکان پێویستن
3. **میتاداتا و وشەسازییەکان** کە بۆ مامەڵە کردن لەگەڵ پاراستنی سەرچاوەکان بە کار دێن، بەتایبەتی بەردەوامی (sustainability)، بەردەستبوون و ئەرشیفکردنی دریژخایەن

پێدراوەی زمان

هەندێک لە گرنگترین جۆرەکانی پێداوە زمانەوانییەکان ئەمانەن:

1. وەسفی زمان وەک ڕێزمان (grammar) و مودێل کردنی فۆرماڵ
2. کۆرپس (corpora / corpus بە کۆ): پێکهاتەیەک لە دەق یان دەنگ بە شێوەی دیجیتاڵ
3. سەرچاوەی واتاناسی-وشەناسی (lexical-semantic): ڕێک خستنی سەرچاوەی زمان بە پێی بۆچوون یان تیۆرییەک لە واتاناسی یان وشەناسیدا
4. ئۆنتۆلۆژی (ontology)، زاراوەناسی (terminology) و فەرهەنگۆک (glossary)
5. مۆدیلی زمان (language model) کە لە سەر کۆرپسە گەورەکاندا فێردەکرێ
6. گرافی زانیاری (knowledge graph) و بنکەی زانیاری (knowledge base)

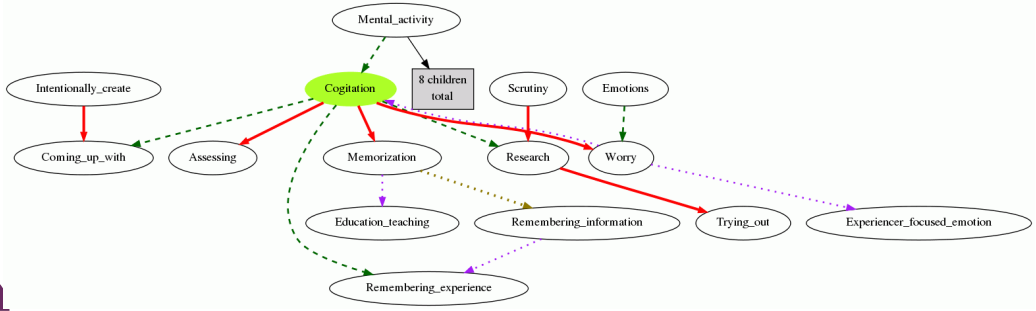
- ڕێزمانی زمان وەک ڕێزمانی کوردیی تاکستۆن [2]
- فۆرمالّیسم و یاساکان

$$\begin{bmatrix} /XY(\text{in})/_V \\ x\,(=\text{چاوگ}) \\ \text{کردن}\ (kirdin) \end{bmatrix} \longleftrightarrow \begin{bmatrix} /XY/_V \\ \text{ڕەگی ڕابردووی}\ x \\ \text{کرد}\ (kird) \end{bmatrix} \longleftrightarrow \begin{bmatrix} /XZ/_V \\ \text{ڕەگی ئێستای}\ x \\ \text{کە}\ (ke) \end{bmatrix}$$

کە تێیدا X ڕەگی کردارەکەیە و Y و Z یش هەردووکیان پاشگری داڕێژانن.

پێداوەی زمان: سەرچاوەی واتاناسی-وشەناسی

- لە سەر بنەمایەکی واتاناسی یان وشەناسیدا بە شێوەیەکی **وەسفکەر** و **پێکهاتەدار** ڕێک خراون.
- هەندێک لە گرینگترینی ئەو سەرچاوەیانە ئەمانەن:
  - فەرهەنگی وشە وەک هەنبانە بۆرینە یان Ferhenga Birûskî [3]
  - وۆردنێت (WordNet) [4] (بەستەر) وەک کوردنێت [5]
  - FrameNet [6]



سەرچاوە زمانەوانییەکان: زاراوەسازی

- زاراوە یان زار (terminology) بە وشەیەک دەکوترێ کە تایبەت بە بوارێکە وەک زاراوەی پزیشکی یان زاراوەی تەکنۆلۆژیا
- سەرجەم زارەکان زۆر جار پێیان دەکوترێ فەرهەنگۆک (glossary)
- هاوشێوەی فەرهەنگ لە سەروشە پێک هاتووە و هەندێک جار زانیاریی زمانەوانیشی لەگەڵدایە. تێرمینۆلۆژیی یەکیەتیی ئەوروپا (بەستەر)
- ئامانجی زاراوە ڕێک خستن و **پۆلاندنی** زانیارییە بۆ ساز کردنی پۆلێنەناسی (taxonomy)
- لە تەکنۆلۆژیای زمان زۆر بەسوودن بۆ زۆرکردنەوەی ئاستی ناسینەوەی وشەکان
- هەندێک بنکەی زاری کوردیی بەسوود
  - وشەنامەی هۆژین: https://hojan.org/dic/
  - یاپراخ: https://dict.linux.krd/

کۆرپس

کۆمەڵێک پێدراوەی زمانەوانی کە بە شێوەی دەق یان دەنگ کۆدەکرێتەوە و وەکوو خاڵی دەستپێک بۆ سەلماندنی گریمانەکان سەبارەت بە زمانێک بە کار دەهێندرێ.

- زۆربەی کۆرپسەکان لانی کەم چەند ملیۆن وشەیان تێدایە.
- کۆرپس تەنیا چیژێک لە زمانەکەیە، نەک زمانەکە بۆ خۆی.
- بە ڕای چۆمسکی، کۆرپس قەت ناتوانێ هەموو دیاردەکانی زمانێک بێنێتە سەلماندن [7].
- بۆ هەر بابەتێک، وەک زانست یان ڕامیاری، و هەر پێناسەیەکی زمان، وەک بنزاراوە یان هەڕەتی قسە کردن، دەتوانین کۆرپسمان هەبێ.
- کۆرپس دەبێ نمونە، هاوسەنگ، نوێنەر و سروشتی بێ [8].
- مژاری سەرەکیی «زمانەوانیی کۆرپس» (corpus linguistics) تێگەیشتن لە کۆرپسە.
- کۆرپسەکان بە بەربڵاوی لە پرۆسەس کردنی زمانیشدا بە کار دەهێندرێن.

سەرچاوە زمانەوانییەکان: کۆرپس (دریژە)

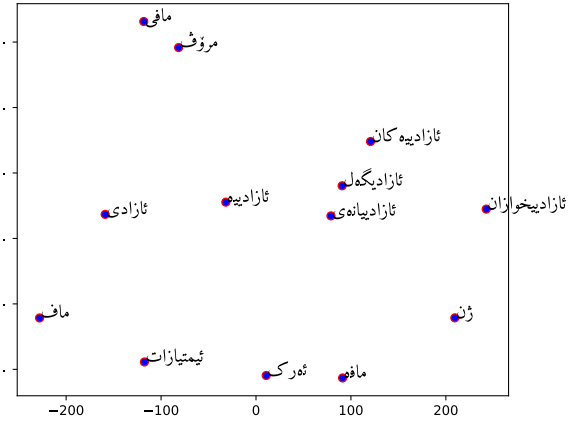
سەرچاوە زمانەوانییەکان: کۆرپس (دریژە)

نێوی کۆرپسەکە	سەرچاوە	بنزاراوە	قەبارە (وشە #)
کۆرپسی پێوان [9]	هەواڵ	سۆرانی	18M
		کورمانجی	4M
کۆرپسی دەقی فۆلکلۆر [10]	گۆرانی، بەند و بەیت	سۆرانی	49K
کۆرپسی ئاسۆسافت [11]	هەواڵ	سۆرانی	80M
کۆرپسی زازا-گۆرانی [12]	هەواڵ	زازاکی	10M
		گۆرانی	2M
کۆرپسی پەروەردە [13]	کتێبی پەروەدە	سۆرانی	0.6M
کۆرپسی مەنچستر [14]	کۆکردنەوەی مەیدانی	سۆرانی	
		کورمانجی	

لیستی هەموو کۆرپسەکان بۆ کوردی لێرەدایە:
https://github.com/sinaahmadi/awesome-kurdish

سەرچاوە زمانەوانییەکان: مۆدیلی زمان

- مۆدیل کردنی ئاماریی زمان کە تایبەتمەندییەکانی زمانێک «فێر دەبێ» ڕۆڵی مۆرفۆلۆژی و سینتاکس، پەیوەندیی نێوان وشەکان و هتد
- بەسوود لە مۆدیل کردنی زۆر دیاردەی زمان و بەربڵاوە لە ئەورۆی تەکنۆلۆژیای زماندا ئەو وشانەی لە باری واتاوە لە وشەی «ماف» نزیکایەتییان هەیە ↓



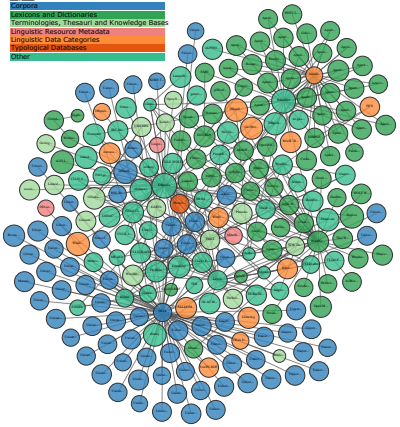
ئۆنتۆلۆژی (ontology)

بە پێکهاتەی ئەو یاسا و زانیارییانەی دەکوترێ کە بۆ ناساندنی چەمکەکانی بواریک و پەیوەندیی نێوانیان بە شێوەی فۆرماڵ بەکاردەهێندرێن.

- ئۆنتۆلۆژییەکان بە توندی فەرمین و بە زمانێکی مەنتیقی دیاری کراون
- «مانگ» یەک وشەیە لە فەرهەنگێکدا، بەڵام دەتوانێ بە چەند شێوە پێناسە بکرێ: مانگی هەتاوی و مانگی هیجری
- سەرچاوە زمانەوانییەکان بە پێی ئاستی فۆرماڵ بوونیان لە ئۆنتۆلۆژی دەچن.
- ئۆنتۆلۆژی لە بواری مەنتیق و کۆمپیوتەردا جیاوازی هەیە. چاو بکە لە [15]
- ئەندازیاریی ئۆنتۆلۆژی بابەتی سەرەکیی بواری «وێبی سێمانتیک» (semantic web) ە.

سەرچاوە زمانەوانییەکان: ئۆنتۆلۆژی (دریژە)

- وشەسازی و مۆدیلێکی پێداواە بۆ سەرچاوەکانی وشە، وەک فەرهەنگەکان [16]
- مۆدیلێکی پێداواە بۆ نیشان دانی پەیوەندی واتایی و بنەماڵەیی نێوان وشەکان [17]
- هەوری سەرچاوە پەیوەندیدارەکان (بەستەر)



سەرچاوە زمانەوانییەکان: گرافی زانیاری و بنکەی زانیاری

- نواندنی زانیاری و پێدراوەکان بە شێوازی گراف (graph)
- پشت بەستن بە بناغەی وێبی سێمانتیک و مۆدیل کردنی پێدراوەکان
- لێک کارکردنی یەکجار زۆری زانیاری لە سەر وێبدا ← چارەسەرێکی باش بۆ ئێستای «پێدراوە پەیوەندیدار»ەکان (linked data)
- بەسوودیشە بۆ زانیاریی زمانەوانی
  ئەژماری وشە ئینگلیسییەکان بە پێی یەکەم پیتیان (بەستەر)
- کۆمەڵی گرافەکانی زانیاری دەبێ بە بنکەی زانیاری (knowledge base)
- لیستی پشکنەرەکانی ئاسمان بە یەک کرتە (بەستەر)
- بنکە بەسوودەکان:
  - Wikidata (سەیری کوردی بکە: بەستەر)
  - DBpedia (سەیری کوردی بکە: بەستەر)
  - ConceptNet (سەیری کوردی بکە: بەستەر)

پێداوەکان بەپێی فۆرماتەکان

```
SEROKÊ IRAQÊ ŞPÊWAZÎYA ENDAMÊ SEROKATÎYA ENCÛMENA
       BILINDA ISLAMÎYA IRAQÊ KIR [10:30] 2010/Oct
       /25 PNA Serokê komara Iraqê Celal Talebanî,
        şpêwazîya endamê serokatîya Encûmena
       Bilinda Islamîya Iraqê Adil Ebdulmehdî û
       şandeya pê re kir. Herdu alîyan daxwaz
       kirin, ku divê hukûmeta nû ya Iraqê zû pêk
       bihê da ku pirs û şarêeyên hevwelatîyan
       çareser bibin û kar û barên gel bi rê ve
       biçin. Di evê dîdarê de ya ku doh êvarê hat
        sazdan, herdu alîyan bîr û nerîn di bareya
        pêkanîna hukûmeta nû ya Iraqê de bi hev
       guherandin û behsa encamên diyaloga navbera
        alîyên siyasî kirin.
```

- ناسراوترین پێناسەی فۆرمات: .txt
- ✓ ساکار، خێرا و ئاسان
- × پێدراوەی بەبێ داڕێژران یان وەسف

```
<entry>
    <form>
        <orth>bend</orth>
        <pron>bænd</pron>
    </form>
    <gramGrp>
        <pos>n</pos>
        <gen>f</gen>
    </gramGrp>
    <sense n="1">
        <cit type="translation" xml:lang="en">
            <quote>bond</quote>
        </cit>
        <cit type="example">
            <quote>divê em êdî li benda sibehê
                  ranewestin.</quote>
            <cit type="translation" xml:lang="en">
                <quote>we shouldn't stand around
                      waiting for tomorrow.</quote>
            </cit>
        </cit>
    </sense>
</entry>
```

- زمانی مارکاپی دریژەپێدەر
  Extensible Markup Language (.xml)
- زمانی پڕۆگرامسازیی مارکاپ هەندێک
  هێمای تایبەت بۆ پێکهێنانی بەڵگە بەکاردێنێ
- ✓ خۆوەسفکەر و داڕێژراو
- ✓ باش بۆ گواستنەوە و دەست وەردان لە پێدراوە
- × هەندێک جار ناخێرا و دریژ
- ✓ پێشنیاری W3C (بەستەر)

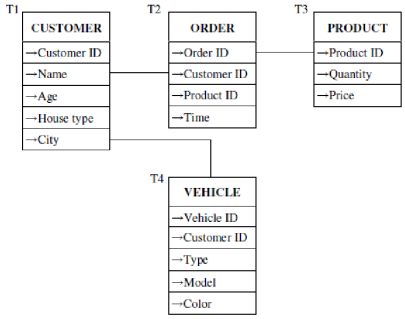
```
{
"firstName": "John",
"lastName": "Smith",
"isAlive": true,
"age": 27,
"address": {
  "streetAddress": "21 2nd Street",
  "city": "New York",
  "state": "NY",
  "postalCode": "10021-3100"
},
"phoneNumbers": [
{
  "type": "home",
  "number": "212 555-1234"
}
}
```

- JavaScript Object Notation (.json)
- ✓ خۆوەسفکەر و داڕێژراو
- ✓ باش بۆ گواستنەوە و دەست وەردان لە پێدراوە
- × کۆل و کورتتر لە XML
- JSON و XML بڕبڕەی

هەڵگواستنەوەی پێدراوەن لە سەر وێبدا.

بنکەدراوە (database)

سیستەمی نەرمەکاڵا بۆ پێناسە کردن، چێ کردن، پاراستن و دەست وەردان لە پێدراوە بەکار دەهێندرێ [18].



- یەکجار خێرا و جێ متمانەیە
- ✓ زۆر بەربڵاو لە ئەورۆی وێبدا
- × خۆوەسفکەر نین و پێکهاتەکان بەپێی مەنتقی پەیوەندییەوە لە خشتەدا ڕێک دەخرێن
- بەنێوبانگترین بنکەدراوەکان پەیوەندیدارن (relational) MySQL، PostgreSQL و Oracle Database

```
:lex_kmr_0278943883 a ontolex:
    LexicalEntry, ontolex:Word ;
  ontolex:canonicalForm :
    form_kmr_0278943883 ;
  rdfs:label "bend"@kmr-latn .

:form_kmr_0278943883 a ontolex:Form ;
  dct:language <www.lexvo.org/page/
    iso639-3/kmr> ;
  ontolex:writtenRep "bend"@kmr-latn ;
  lexinfo:partOfSpeech lexinfo:noun ;
  lexinfo:gender lexinfo:feminine ;
  lexinfo:number lexinfo:singular ;
  ontolex:sense :kmr_9120343779_sense.
```

- چوارچێوەی وەسفکردنی سەرچاوە(Resource Description Framework (.ttl)
- ✓ خۆوەسفکەر و داڕێژراو
- ✓ زمانی وێب و پێدراوە پەیوەندیدارەکان
- ✓ خێرا و بەربڵاو
- × هەندێک جار ئەندازەی پێدراوەکان زۆر دەکاتەوە
- ✓ پێشنیاری W3C (بەستەر)
- لێرەدا دەتوانن لەگەڵ RDFدا زۆرتر ئاشنا بن: بەستەر.

بەکارهێنانی سەرچاوە زمانییەکان

- هاوشێوەی بنەماکانی بەڕێوەبردنی پێداوە زانستییەکان [19]، سەرچاوەیەکی زمانەوانی باش دەبێ:
  - **F**indable: بدۆزرێتەوە
  - **A**ccessible: بەردەست بێ
  - **I**nteroperable: کاری لەگەڵ بکرێ
  - **R**eusable: بە کەڵک بێتەوە

- گرینگایەتیی پێداوە زمانەوانییە پەیوەندیدارەکان (linguistic linked data) [20]
- پێداوە بناغەی تێگەیشتن لە زمان و گرینگایەتیشیان لە زانستە مرۆییە دیجیتاڵییەکاندا (digital humanities) هەیە.

بەشی پراکتیک

ئەم بەشە لە خولەکە، بەشێکی پراکتیکیشی هەیە کە خوێندکار دەتوانێ لە ڕێگای ماڵپەڕی خولەکەوە شوێنی بکا.

[1] Unsupervised cross-lingual representation learning.
`https://ruder.io/unsupervised-cross-lingual-learning`.
Accessed: 2022-07-30.

[2] Wheeler M. Thackston.
Sorani Kurdish–A Reference Grammar with Selected Readings.
Harvard University, 2006.

[3] Michael L Chyet and Martin Schwartz.
Kurdish-English Dictionary (Kurmanji).
Yale University Press, 2003.

[4] George Miller, Christiane Fellbaum, Judy Kegl, and Katherine Miller.
Wordnet: An electronic lexical reference system based on theories of lexical memory.
Revue quebecoise de linguistique, 17(2):181–212, 1988.

[5] Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili.
Towards building Kurdnet, the Kurdish Wordnet.
In *Proceedings of the Seventh Global Wordnet Conference*, pages 1–6, 2014.

[6] Collin F Baker, Charles J Fillmore, and John B Lowe.
The Berkeley Framenet project.
In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90.
Association for Computational Linguistics, 1998.

[7] Jacqueline Léon.
Claimed and unclaimed sources of corpus linguistics.
Henry Sweet Society for the History of Linguistic Ideas Bulletin, 44(1):36–50, 2005.

[8] Guillaume Desagulier, Guillaume Desagulier, and Amboy.
Corpus linguistics and statistics with R.
Springer, 2017.

[9] Kyumars Sheykh Esmaili, Donya Eliassi, Shahin Salavati, Purya Aliabadi, Asrin Mohammadi, Somayeh Yosefi, and Shownem Hakimi.
Building a test collection for Sorani Kurdish.
In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE, 2013.

[10] Sina Ahmadi, Hossein Hassani, and Kamaladdin Abedi.
A Corpus of the Sorani Kurdish Folkloric Lyrics.
In *Proceedings of the 1st Joint Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) Workshop at the 12th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France, 2020.

[11] Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini.
Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus.
Digital Scholarship in the Humanities, 35(1):176–193, 2020.

[12] Sina Ahmadi.
Building a corpus for the zaza–gorani language family.
In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2020)*, 2020.

[13] Roshna Omer Abdulrahman, Hossein Hassani, and Sina Ahmadi.
A Rule-Based Kurdish Text Transliteration System.
2019.

[14] Yaron Matras.
Revisiting Kurdish dialect geography: findings from the Manchester Database.
Current issues in Kurdish linguistics, 1:225, 2019.

[15] Nicola Guarino and Pierdaniele Giaretta.
Ontologies and knowledge bases.
Towards very large knowledge bases, pages 1–2, 1995.

[16] John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano.
The ontolex-lemon model: development and applications.
In *Proceedings of eLex 2017 conference*, pages 19–21, 2017.

[17] Mustafa Jarrar.
The Arabic ontology–an Arabic wordnet with ontologically clean content.
Applied ontology, (Preprint):1–26, 2021.

[18] Thomas M Connolly and Carolyn E Begg.
Database systems: a practical approach to design, implementation, and management.
Pearson Education, 2005.

[19] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al.
The fair guiding principles for scientific data management and stewardship.
Scientific data, 3(1):1–9, 2016.

[20] Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Max Ionov, Penny Labropoulou, Francesco Mambrini, John McCrae, Émilie Pagé-Perron, Marco Passarotti, Salvador Ros, and Ciprian-Octavian Truica.
When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data.
http://www.semantic-web-journal.net/content/when-linguistics-meets-web-technologies-recent-advances-modelling-linguistic-linked-open-0, 2021.

[21] Steven Bird and Gary Simons.
Extending dublin core metadata to support the description and discovery of language resources.
Computers and the Humanities, 37(4):375–388, 2003.