Interconnecting lexicographic resources.
In search for a model

Dan Cristea
“Alexandru Ioan Cuza” University of Iași
Institute of Computer Science of the
Romanian Academy

dcristea@info.uaic.ro
Topics

• Why would one want to connect linguistic resources?
• Parameterising the needs
• Standardisation helps interconnecting
• A bunch of notorious resources
• How would this work?
• Final remarks
Why would one want to connect linguistic resources?

- Use case 1: 100 Romanian dictionaries aligned
  - **CLRE. Essential Romanian Lexicographical Corpus.**
  - 100 dictionaries aligned at entry and, partially, sense levels (2010–2013, at Institute A.Philippide of the Romanian Academy, in Iași)
  - dictionaries’ list at:
    - http://85.122.23.90/resurse/Lista-dictionarelor.doc
  - written in 3 types of alphabets: Cyrillic, transition and Latin
  - large diversity of formatting styles
The CLRE project

Bucharest, 14-15 December 2012

COST-ENeL, Bled, 29-30 September 2014
DARĂPĂSTORE, o. f. Construcții de călătorie, căi de drum, mătase, drumuri, drumuri de muncă, drumuri de soare, drumuri de zăpadă, drumuri de zăpadă.
The CLRE project

Bucharest, 14-15 December 2012

COST-ENeL, Bled, 29-30 September 2014
Dictionar de informatică

The CLRE project

Bucharest, 14-15 December 2012

COST-ENeL, Bled, 29-30 September 2014
Processing in CLRE

- Scanning
- OCR – Abby Fine Reader 9
- Parsing entries => XML
- Manual verification
- Indexing and alignment
CLRE – manual verification

45

✓ AMERICANIZÁ

✓ AMBULĂCĂRI, -A, ambulațion, -e adj. (fr. ambulancier) Cu privire la ambulația, de ambulație.

✓ AMBULĂCIE, -ă, ambulație s.n. (fr. ambulance) Tub subtotal situat pe fața inferioară a corpului echipat în diferite forme și termeni pe care o vorbim, care se referă la locomotiva, coajă și pieță.

✓ AMBULANT, -A, ambulant, -ei adj. (fr. ambulant; lat. ambulans; mtl) Care se deplasează dintr-un loc în altul, care nu are un loc de reședință fixă.

✓ AMBULANTA, ambulanta s.f. (fr. ambulance) Mă loc de transport al bolnavilor, al ranitilor etc.; s-a variat (2).


✓ AMBUSCADA, ambuscada s.f. (fr. ambuscade) Mâna nume de atac prin surprinzare asupra unui inamic în miocare.

✓ AMBUSCAT, -Ă, ambuscadă -ă adj., s.m. s.f. (fr. ambuscade) Echilibrul scutil de obligațiile peticioase din timpul razboiului. ✓ AMEȘUGURA, ambești s.f. (fr. ambouchure) Porta a unui instrument medical prin care se susține acel su gura.

✓ AMBUTEA, ambutea s.f. (fr. emboutie) A bloc ca o joc (futură, maritime etc.) cu vehicule, nave etc.

✓ AMBUTEA, ambutea s.f. (fr. emboutie) 1. Imbuterie. 2. Blocaj de circulație rutieră sau navală din cauza abrumoarei.

✓ AMBUTIESĂ, ambutează s.f. (fr. emboutissant) A supune un metat unor secvențe mecanice pentru a-i da o anumită formă.

Iași, 25-26 September 2013
Why would one want to connect linguistic resources?

• Use case 2: align WN with an explanatory dictionary
  – a WN synset:
    \[ \text{pos} \,(\text{def, ex, } w_1^{s_1} \ldots w_k^{s_k} \ldots w_n^{s_n}) \]
  – an explanatory dictionary entry:
    \[ w_k, \text{pos}, <w_k^{s_1}, \text{def}_1, \text{ex}_1> \ldots <w_k^{s_k}, \text{def}_k, \text{ex}_k> \ldots <w_k^{s_m}, \text{def}_m, \text{ex}_m> \]
Why would one want to connect linguistic resources?

• Use case 2: align WN with an explanatory dictionary
  – synsets of a word $w_k$, pos:
    
    
    (def$_1$, ex$_1$, ...$w_k^{s1}$ ...

    ... 

    (def$_k$, ex$_k$, ...$w_k^{sk}$ ...

    ... 

    (def$_m$, ex$_m$, ...$w_k^{sm}$ ...
  
  – the explanatory dictionary entry of the word $w_k$, pos:
    
    $w_k$, pos, <$w_k^{s1}$,def$_1$,ex$_1$>,... <$w_k^{sk}$,def$_k$,ex$_k$>,... <$w_k^{sm}$,def$_m$,ex$_m$>
Why would one want to connect linguistic resources?

- Use case 2: align WN with an explanatory dictionary
  - a WN synset:
    \[ \text{pos} \quad (\text{def, ex, } w_1^{s_1} \ldots w_k^{s_k} \ldots w_n^{s_n}) \]
  - explanatory dictionary entries:
    \[ w_1, \text{pos}, \ldots <w_1^{s_1}, \text{def, ex}> \ldots \]
    \[ w_k, \text{pos}, \ldots <w_k^{s_k}, \text{def, ex}> \ldots \]
    \[ w_n, \text{pos}, \ldots <w_n^{s_n}, \text{def, ex}> \ldots \]

COST-ENeL, Bled, 29-30 September 2014
Why would one want to connect linguistic resources?

- Use case 3: the TOT problem or the forgotten word
The first thought: standardisation

• Lexical Markup Framework (LMF)
  – What is it?
    • a common model for creation and use of lexical resources
  – With what goal?
    • to manage the exchange of data between and among these resources
    • to enable the merging of a large number of individual electronic resources to form extensive global electronic resources
Near-standard

• Text Encoding Initiative (TEI)
  – What is it?
    • an inventory of the features most often deployed for computer-based text processing
    • recommendations about suitable ways of representing these features
  – With what goal?
    • to facilitate processing by computer programs
    • to facilitate the loss-free interchange of data amongst individuals and research groups using different programs, computer systems, or application software
Standardisation

- Text Encoding Initiative (TEI)
  - Example of a dictionary entry serialisation
    (from TEI Guidelines)

\[
\text{disproof (dIs"pru:f) n. 1. facts that disprove something. 2. the act of disproving. CED}
\]

\[
<\text{entry}>
  <\text{form}>
    <\text{orth}>\text{disproof}</\text{orth}>
    <\text{pron}>dls"pru:f</\text{pron}>
  </\text{form}>
  <\text{gramGrp}>
    <\text{pos}>n</\text{pos}>
  </\text{gramGrp}>
  <\text{sense n="1"}>
    <\text{def}>facts that disprove something.</\text{def}>
  </\text{sense}>
  <\text{sense n="2"}>
    <\text{def}>the act of disproving.</\text{def}>
  </\text{sense}>
</\text{entry}>
\]

COST-ENeL, Bled, 29-30 September 2014
LMF and TEI content and... discontent

• The TEI format may be used as an interchange format, permitting sharing of resources even when their local encoding schemes differ.
• Both LMF and TEI model lexical material at a deep representational detail...
LMF and TEI
content and... discontent

• TEI intention:
  – guidance for individual or local practice in text creation and data capture
  – support of data interchange
  – support of application-independent local processing

• Opening good possibilities of querying

• But how would function the interconnection?...
Parameterising the needs

• If I want to connect two resources, simply merge the contents
• Then be able to interrogate the merged resource by taking advantage of peculiarities in each resource
Parameterising the needs

• Able to represent variations in word forms, alternate orthography, diachronic morphology
• Easy navigation by applying various filtering criteria

COST-ENeL, Bled, 29-30 September 2014
Parameterising the needs

• Very often lexicographic data is hierarchical
  – for instance, a sense of a dictionary entry contains a definition, examples, but also sub-senses

• Organise even recursive searches
  – give me the definition neighbouring sphere of depth 2 of the word captain (take all senses of the entry captain and form the list of words in the corresponding definitions, then for each of them take all their senses and collect again words in their definitions)
The idea

• Representing lexical information as feature structures centred on word’s lemmas

– disproof (dI$s"pru:f) n. 1. facts that disprove something. 2. the act of disproving. CED

[lemma=disproof, entry=[pron=dI$s"pru:f, pos=n, sense=[n=1, def=facts that disprove something], sense=[n=2, def=the act of disproving], res=CED]]

COST-ENeL, Bled, 29-30 September 2014
Representing lexical entries as feature structures

- **lemma**: disproof
- **entry**:
  - **pron**: dls"pru:f
  - **pos**: n
  - **sense**:
    - **n=1**: def=facts that disprove something
    - **n=2**: def=the act of disproving
  - **sense**:
  - **res**: CED

Cambridge English Dictionary

COST-ENeL, Bled, 29-30 September 2014
Representing lexical entries as feature structures

- Graph representation

CED

COST-ENeL, Bled, 29-30 September 2014
Representing lexical entries as feature structures

**Lemma:** disproof

**Entry:**

- **Pronunciation:** dls"pru:f
- **Part of Speech:** n
- **Sense 1:**
  - Definition: the action of disproving
- **Sense 2:**
  - Definition: evidence that disproves
- **Resource:** MWCD

Merriam-Webster's Collegiate Dictionary

COST-ENeL, Bled, 29-30 September 2014
Representing lexical entries as feature structures

- Graph representation

MWCD

COST-ENeL, Bled, 29-30 September 2014
How could lexical entries be merged?

- Entries of the same word from different dictionaries

CED

MWCD
Merging lexical entries

• Distinct parts

disproof
dIs"pru:f
lemma
pron
entry
pos
n
sense
n
sense
n
res
def
1
facts that...
sense
def
2
the act of...

COST-ENeL, Bled, 29-30 September 2014
Merging lexical entries

disproof
lemma
dls"pru:f
pron
entry
pos
n
X (new)

sense
n
def
res
MWCD

sense
n
def
res
CED

sense
n
def
facts that...

sense
n
def
the act of...

sense
n
def
evidence that...

the action of...

COST-ENeL, Bled, 29-30 September 2014
COST-ENeL, Bled, 29-30 September 2014
The WordNet search for *disproof*

**Noun**

- **S: (n)** *disproof, falsification, refutation* (any evidence that helps to establish the falsity of something)
- **S: (n)** *falsification, falsifying, disproof, refutation, refutation* (the act of determining that something is false)
Feature structures representation for the WN synsets of *disproof*

- disproof
  - lemma
  - synsets
  - n
  - pos
  - synset
  - synset
  - synset
  - lex
  - gloss
  - gloss

(*, falsification, refutation) any evidence that helps to establish the falsity of something

(falsification, falsifying, *, refutation, refutal) (the act of determining that something is false

COST-ENeL, Bled, 29-30 September 2014
The WordNet search for *discount*

**Noun**
- *S:* (n) discount, price reduction, deduction (the act of reducing the selling price of merchandise)
- *S:* (n) discount rate, discount, bank discount (interest on an annual basis deducted in advance on a loan)
- *S:* (n) rebate, discount (a refund of some fraction of the amount paid)
- *S:* (n) deduction, discount (an amount or percentage deducted)

**Verb**
- *S:* (v) dismiss, disregard, brush aside, brush off, discount, push aside, ignore (bar from attention or consideration) "She dismissed his advances"
- *S:* (v) discount (give a reduction in price on) "I never discount these books—they sell like hot cakes"
Representing WordNet synsets

discount

lemma

cost-ENeL, Bled, 29-30 September 2014
How could dictionary entries be merged with WN synsets?

- disproof
- lemma
- entry
- pron
- pos
- n
- sense
- synsets
- synset
- lex
- gloss
- any evidence that helps to establish the falsity of something
- (falsification, falsifying, *, refutation, refutal)
- (the act of determining that something is false)
- CED
- n
- def
- facts that disproves smth
- the act of disproving
- (*, falsification, refutation)
- COST-ENeL, Bled, 29-30 September 2014
Merging a dictionary entry with a WN entry

- **Lemma**: disproof
- **Entry**: dls"pru:f
  - **Position**: n
  - **Sense**: 1. facts that disproves smth
    - **Definition**: the act of disproving
  - **Sense**: 2. the act of disproving
  - **Res**: CED
  - **Synsets**: (*, falsification, refutation)
    - **Lex**: any evidence that helps to establish the falsity of something
      - **Lex**: (falsification, falsifying, *, refutation, refutal)
      - **Gloss**: (the act of determining that something is false)

COST-ENeL, Bled, 29-30 September 2014
Merging a dictionary entry with a WN entry

- **disproof**
  - **entry**
    - **sense**
      - **n**
        - **facts that disproves smth**
  - **the act of disproving**
  - **CED**

- **synsets**
  - **synset**
    - **gloss**
      - any evidence that helps to establish the falsity of something
      - (falsification, falsifying, *, refutation, refutal)

- **COST-ENeL, Bled, 29-30 September 2014**
Going one step further

- Feature structures are hierarchical data
- Codd: hierarchical data can be represented as relational tables
Representing feature structures as relational tables


COST-ENeL, Bled, 29-30 September 2014
Representing feature structures as relational tables


COST-ENeL, Bled, 29-30 September 2014
Representing feature structures as relational tables


COST-ENeL, Bled, 29-30 September 2014
Representing feature structures as relational tables

from

COST-ENeL, Bled, 29-30 September 2014
Representing feature structures as relational tables


COST-ENeL, Bled, 29-30 September 2014
Relational operators

- **Projection**: $\pi_{a_1,\ldots,a_n}(R) \Rightarrow$ a relation containing only values of attributes $a_1,\ldots,a_n$ from the relation $R$

- **Selection**: $\sigma_{\varphi}(R)$, with $\varphi$ is logical condition $\Rightarrow$ only tuples verifying the condition $\varphi$ are retained from the relation (or the set) $R$

- **Join**: $R \bowtie S \Rightarrow$ the set of all attributes in $R$ and $S$ that are equal on their common attributes

- **Union**: $R \cup S \Rightarrow$ a table representing the union of the two relations

COST-ENeL, Bled, 29-30 September 2014
Interrogating a dictionary

- Citations before 1850 of the entry “symphony”.

π_{orth}(σ_{lemma=“symphony” & yr<1850} (WORDENTRYSENSECIT))

COST-ENeL, Bled, 29-30 September 2014
Interrogating a combination between a dictionary and a wordnet

- All synonyms of nouns belonging to citations dated before 1850, sorted lexicographically.
Interrogating a combination between a dictionary and a wordnet

• All synonyms of nouns belonging to citations dated before 1850, sorted lexicographically.

The citations of the title word $w$: 

$$\pi_{\text{orth}}(\sigma_{\text{lemma}=w \land \text{yr}<1850} (\text{WORD} \bowtie \text{ENTRY} \bowtie \text{SENSE} \bowtie \text{CIT}))$$

COST-ENeL, Bled, 29-30 September 2014
Interrogating a combination between a dictionary and a wordnet

• All synonyms of nouns belonging to citations dated before 1850, sorted lexicographically.

Unify and lemmatise words belonging to the citations:
\[ \text{lem}(U(\pi_{\text{orth}}(\sigma_{\text{lemma}=w \& \text{yr}<1850}\ (\text{WORD} \bowtie \text{ENTRY} \bowtie \text{SENSE} \bowtie \text{CIT})))) \]

COST-ENeL, Bled, 29-30 September 2014
Interrogating a combination between a dictionary and a wordnet

- All synonyms of nouns belonging to citations dated before 1850, sorted lexicographically.

\[ \pi_{\text{lex}} \left( \sigma_{\text{pos}=n \land \text{lemma} \in \text{lem}(U(\pi_{\text{orth}}(\sigma_{\text{lemma}=w \land \text{yr}<1850})(\text{WORD} \bowtie \text{ENTRY} \bowtie \text{SENSE} \bowtie \text{CIT})))}(\text{WORD} \bowtie \text{SYNS} \bowtie \text{SYN})) \right) \]

COST-ENeL, Bled, 29-30 September 2014
Conclusions

• I did not propose any model, I simply made some observations (nothing is really new)
• Linking lexicographic resources:
  – one resource => TEI representation => as feature structures => hierarchical graphs => relational tables
  – more resources => unifications of tables
  – use query and relational operators for interrogation
Discussion

• Only a sketch
  – a lot of details should still be filled in
  – the good news: XML structures (the native language of TEI) accept direct representations as database records: XSLT => opening direct access to a complex querying language: XQuery => mimicking the relational operators and adding more facilities
Discussion

• Another good news
  – representing variable depth structures
  – recursive hierarchies: Kamfonas
    • *Fixed depth dimensions are simpler to implement, maintain and query... Hierarchies that have variable depth or an uncertain number of levels... can often benefit if implemented as recursive hierarchies.*
    
    http://www.kamfonas.com/id3.html
Discussion

• Even more good news (hopes)
  – interrogations can be formulated in natural language => an interpreter translates them in the query language of a DBMS system
  – as such, a handy tool at the benefit of lexicographers
Acknowledgements

• Work partially supported by the project *The Computational Representative Corpus of Contemporary Romanian Language*, a project of the Romanian Academy and partially by the COST-ENeL project

• I thank Isabelle Tamba and Mădălin Pătrașcu for the slides describing the CLRE project
Thank you!