Visualisation as an afterthought: lessons learned

Arvi Tavast[1], **Maria Tuulik**[2], **Jelena Kallas**[2]
[1] Qlaara Labs, Tallinn, Estonia
[2] Institute of the Estonian Language, Tallinn, Estonia
arvi@qlaara.com, maria.tuulik@eki.ee, jelena.kallas@eki.ee
Budapest, 24 February 2017

Contents

.

The process so far

Estonian Collocation Dictionary (ECD)

- Monolingual online scholarly dictionary aimed at learners of Estonian as a foreign or second language at the upper intermediate and advanced levels
- Institute of the Estonian Language in collaboration with Lexical Computing Ltd.
- 463M word corpus
- Headwords, collocates and example sentences automatically extracted
- Manual cleaning, 6000 of 10000 headwords completed
- To be published 2018
- Strong paper heritage

The database structure

A sample XML entry from the ECD database (simplified for readability)

```
<m>lahe</m>
<relg>
    <reln>predicate_Adj_translative_of</reln>
    <colg>
        <mse>lahedaks</mse>
        <col>pidama</col>
        <cfr>19</cfr><csc>12.377563</csc>
        <cng>
            <cn>Tol ajal peeti suitsetamist lahedaks, ohutuks ning tervislikuks.</cn>
        </cng>
    </colg>
    <colg>
        <mse>lahedaks</mse>
        <col>tegema</col>
        <cfr>18</cfr><csc>12.324435</csc>
        </cng>
    </colg>
</relg>
<relg>
    <reln>Adj_Vda</reln>
    <colg>
        <mse>lahe</mse>
        <col>vaadata</col>
        <cfr>32</cfr><csc>5.723221</csc>
    </colg>
</relg>
```

The user interface

The same sample entry from the current working version of user interface

lahe omadussõna **18285** :

Tegusõnaga

- predicate_Adj_saav_of **79**
  - lahedaks pidama
  - lahedaks tegema
- predicate_Adj_nimetav_of **35**
  - lahe tunduma

New requirements

added as an afterthought

- Visualisation of the dictionary would be attractive and improve usability.
- Collocation data should be reusable for inclusion in other dictionaries.

The database structure

A sample XML entry from the ECD database (simplified for readability)

```
<m>lahe</m>
<relg>
    <reln>predicate_Adj_translative_of</reln>
    <colg>
        <mse>lahedaks</mse>
        <col>pidama</col>
        <cfr>19</cfr><csc>12.377563</csc>
        <cng>
            <cn>Tol ajal peeti suitsetamist lahedaks, ohutuks ning tervislikuks.</cn>
        </cng>
    </colg>
    <colg>
        <mse>lahedaks</mse>
        <col>tegema</col>
        <cfr>18</cfr><csc>12.324435</csc>
        </cng>
    </colg>
</relg>
<relg>
    <reln>Adj_Vda</reln>
    <colg>
        <mse>lahe</mse>
        <col>vaadata</col>
        <cfr>32</cfr><csc>5.723221</csc>
    </colg>
</relg>
```

Implementational issues

Problematic design decisions from earlier phases

- Representation of nodes and collocates
- Generalisation of context examples
- Missing frequency and salience data

Representation of nodes and collocates

Collocates not necessarily headwords themselves

Exceptions:

- typing errors
- inadvertent omissions
- deliberate decisions to only include a collocation in one direction

Representation of nodes and collocates

Collocates semantically ambiguous

Not known where to connect to:

- Homonyms
- Polysemes

Manual disambiguation?
Omit the link?
Replace deterministic link with a search?

Representation of nodes and collocates

Collocates morphologically ambiguous

- Multiple potential analyses
- No context to disambiguate from
  - (except the headword)

Manual disambiguation?
Omit the link?
Replace deterministic link with a search?
Semantic disambiguation based on ECD itself?

Generalisation of context examples

from individual collocation to type of collocation

Context examples:

- Retrieved from the corpus for each collocation
- Generalised manually to type of collocation
- Stored at the first collocation of the type

Ok on paper, but:

- Separate mechanism for retrieving the example from another collocation
- Automatic corpus linking doesn't make sense

User preference?

Missing frequency and salience data

Deemed unnecessary for the user

Frequency and salience data:

- Used in automatic generation
- Stored in the database for automatically generated collocations
- Missing for manually added entries
- Deemed unnecessary; order is enough for the user

But for the visualisation:

- Data missing
- No obvious algorithm for filling the gaps
- Too much work to restore data

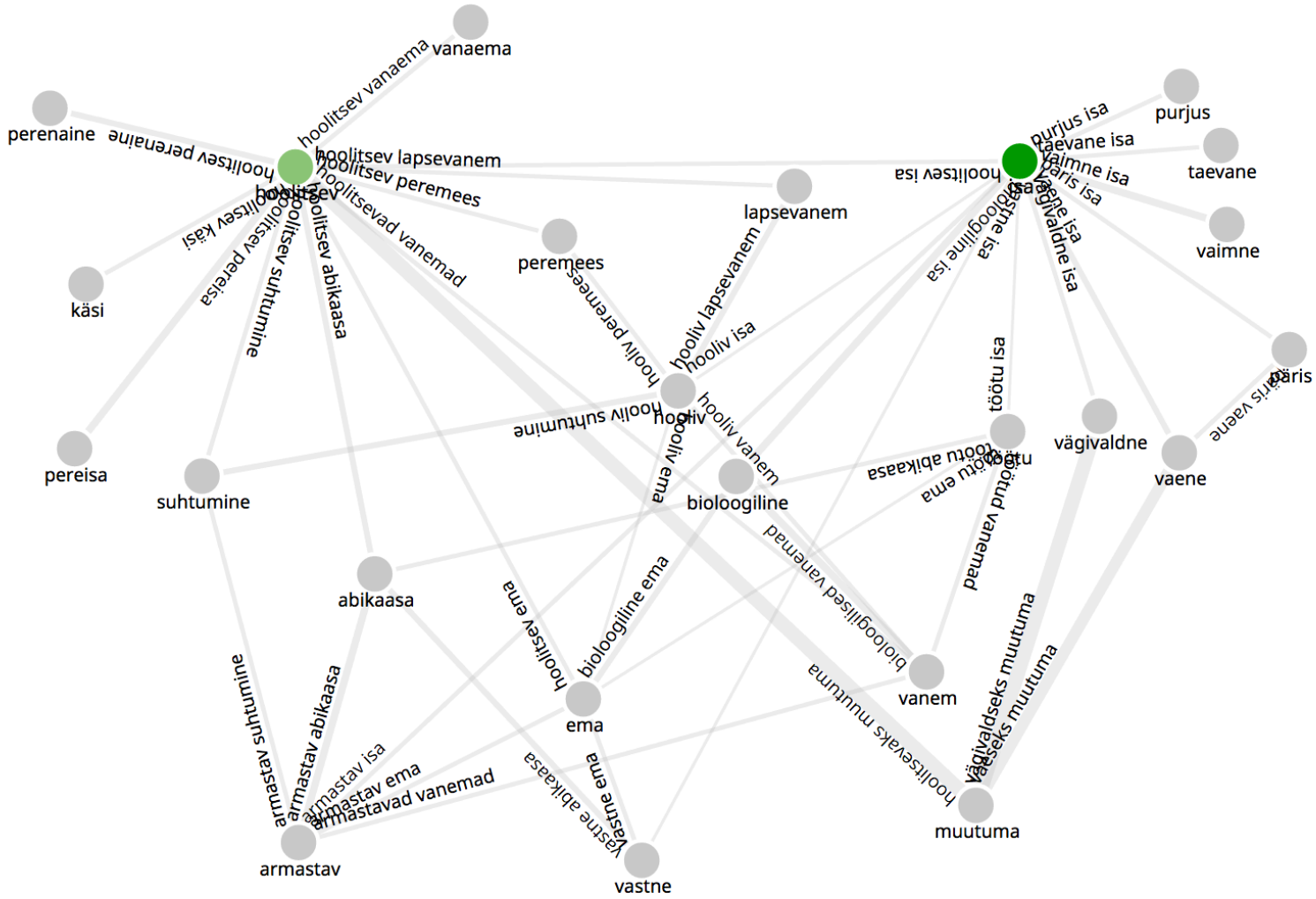
Fundamental issues

Problems that need solving regardless of the implementation

- Collocates in non-canonical forms
- Selection of collocation types
- Symmetry of collocations
- Collocations with more than two members

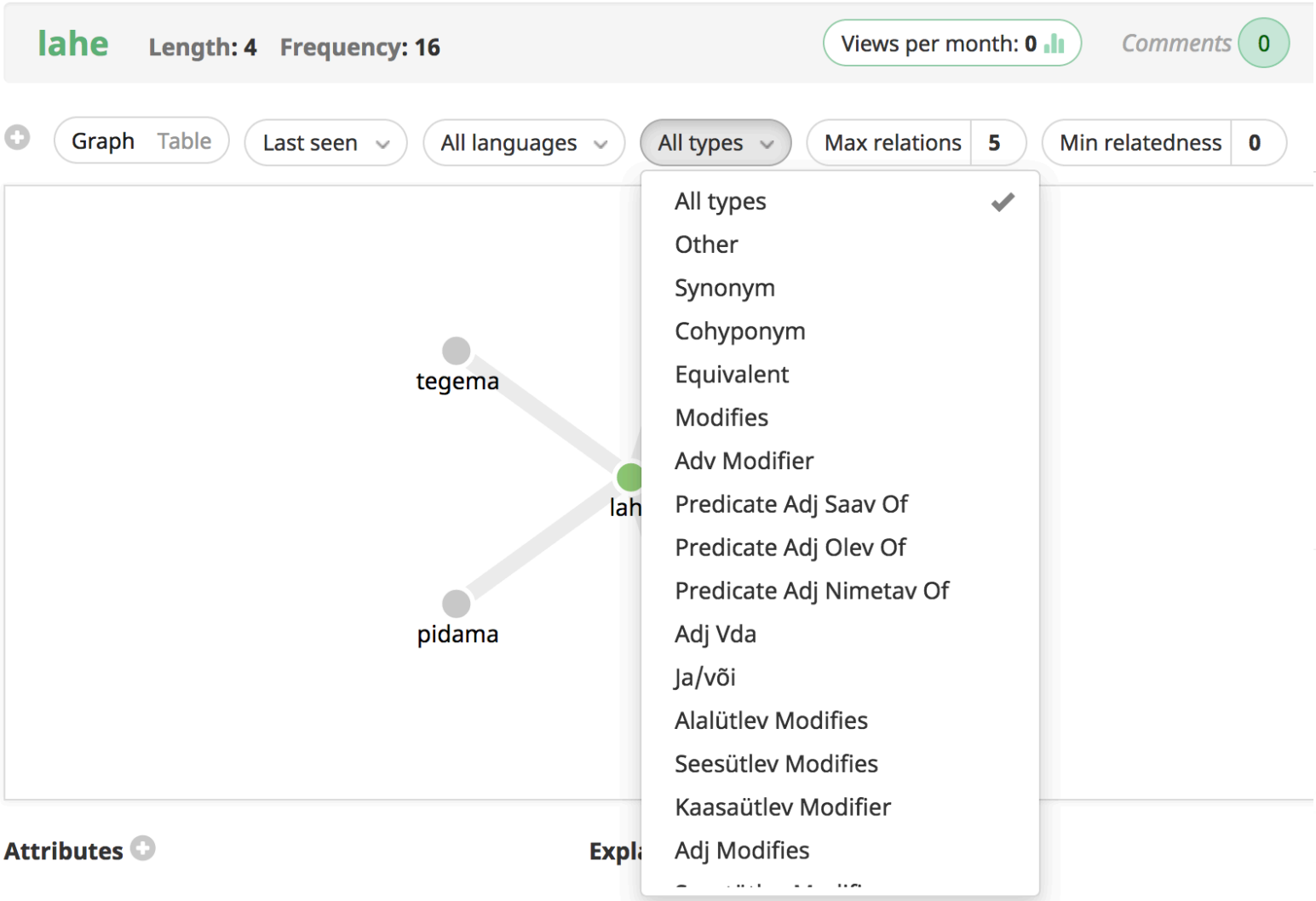
Collocates in non-canonical forms

Not obvious for the user based on lemmata



Selection of collocation types

28 types on two levels: precision *vs* usability



Symmetry of collocations

In the corpus:

- If *A* co-occurs with *B*, then *B* inevitably co-occurs with *A*

In the dictionary, not necessarily:

- Frequency distribution of collocations varies across words
  - like *very* with many adjectives
- User expectation?
- But when navigating a visualisation?

Collocations with more than two members

no obvious way to visualise

- hea välja nägema
- weapon of mass destruction
- chief executive officer

Hierarchical collocations?

- hea (välja nägema)
- weapon of (mass destruction)
- chief executive officer ?

Longer headwords?

Conclusion

.

The collocation dictionary can't be visualised.

Especially not from its current data structure.

But:
Reality is largely negotiable. If you stress-test the boundaries, you quickly discover that most limitations are just a fragile collection of socially reinforced rules that you can choose to break at any time.

–Tim Ferriss, Tools of Titans.