An Unsupervised Approach to Characterize the Adjectival Microstructure in a Hungarian Monolingual Explanatory Dictionary

Enikő Héja[1] Noémi Ligeti-Nagy[1] László Simon[2] Veronika Lipp[2]

[1]Language Technology Research Group
Hungarian Research Centre for Linguistics

[2]Lexical Knowledge Representation Research Group
Hungarian Research Centre for Linguistics

eLex 2023, Brno, 27–29 June

Structure

1. Motivation
2. Problem: when to tell apart meanings in the case of polysemy?
   - Near-synonymy
3. Solution: distributional criteria for meaning distinction
4. Modelling meaning distinction
   - Representation of adjectives, adjectival meanings, polysemies
   - Detecting the subcategorized nouns
5. Method
6. Conclusion and future work

Structure

Motivation – Lexicographic background

The Hungarian Research Centre for Linguistics aims to update the *The Explanatory Dictionary of the Hungarian Language* [EDHL] (Bárczi, G. & Országh, L. eds; 1959–1962)

- Which was based on lexicographers' intuition

Motivation – Lexicographic background

The Hungarian Research Centre for Linguistics aims to update the *The Explanatory Dictionary of the Hungarian Language* [EDHL] (Bárczi, G. & Országh, L. eds; 1959–1962)

- Which was based on lexicographers' intuition
- The main methodological issue of the new version of EDHL is to obtain an objective lexical profile for each dictionary entry

Motivation – Lexicographic background

The Hungarian Research Centre for Linguistics aims to update the *The Explanatory Dictionary of the Hungarian Language* [EDHL] (Bárczi, G. & Országh, L. eds; 1959–1962)

- Which was based on lexicographers' intuition
- The main methodological issue of the new version of EDHL is to obtain an objective lexical profile for each dictionary entry
- By means of corpus-based and/or corpus-driven methods
  - Utilising huge amount of text corpora

Motivation – Lexicographic background

The Hungarian Research Centre for Linguistics aims to update the *The Explanatory Dictionary of the Hungarian Language* [EDHL] (Bárczi, G. & Országh, L. eds; 1959–1962)

- Which was based on lexicographers' intuition
- The main methodological issue of the new version of EDHL is to obtain an objective lexical profile for each dictionary entry
- By means of corpus-based and/or corpus-driven methods
  - Utilising huge amount of text corpora
- **Adjectives** are the focus of our research
  - They are especially difficult to divide into distinct senses (Moon, 1987)
  - They are rather overlooked in the lexical semantic literature
  - Unsupervised word sense induction relying on substantial amount of *unlabeled data*
  - ⇒ Minimal presuppositions about senses and subsenses

Motivation – Consistent Methodology

- The corpus-driven technique provides a more objective conception of polysemic meaning distinction
  - It relies on distributional criteria to tell apart (sub)senses – a novel contribution to the field
  - The adjectival meanings are distilled from cc. 170 million sentences (Nemeskey, 2020)
  - Contextual information is retrieved from the 180-million word HNC (Váradi, 2002)
- Can be easily modelled by a graph-based approach
- Expectation: the collaboration between lexicographers and NLP researchers results in:
  1. an improved WSI methodology and
  2. the development of data-oriented, explicit lexicographic editing principles that apply to both the macrostructure and microstructure of the dictionary.

Structure

1. Motivation
2. Problem: when to tell apart meanings in the case of polysemy?
   - Near-synonymy
3. Solution: distributional criteria for meaning distinction
4. Modelling meaning distinction
   - Representation of adjectives, adjectival meanings, polysemies
   - Detecting the subcategorized nouns
5. Method
6. Conclusion and future work

What is polysemy?

- Polysemy: "multiple meanings that are somehow related to each other" (Ježek 2016)
- ⇒ What is a meaning?
- WordNet-based approach: meanings are constituted by sets of synonyms (synsets)
- Words with multiple meanings belong to multiple synsets
- Synonymy: "iff two expressions are interchangeable in **every context** preserving the original meaning"
  ⇒ too strong

Near-synonymy

- Instead of synonymy: two expressions are near-synonyms iff they are interchangeable in a restricted set of contexts without changing the meaning (cf. Ploux & Victorri, 1998)

Examples

finom 'fine' and *lágy* 'soft' are synonyms before nouns related to MUSIC (eg. *zene* 'music', *ritmus* 'rhythm', *dallam* 'melody')

- An adjective is considered to have multiples meanings if it belongs to multiple near-snyonymy classes

Structure

1. Motivation
2. Problem: when to tell apart meanings in the case of polysemy?
   - Near-synonymy
3. Solution: distributional criteria for meaning distinction
4. Modelling meaning distinction
   - Representation of adjectives, adjectival meanings, polysemies
   - Detecting the subcategorized nouns
5. Method
6. Conclusion and future work

Four Distributional Criteria for Meaning Distinction

1. There is (at least) one near-synonym for each sense of the adjective.

Four Distributional Criteria for Meaning Distinction

1. There is (at least) one near-synonym for each sense of the adjective.
2. There is a set of context nouns that form grammatical constructions both with the original adjective and with the near-synonym.

Four Distributional Criteria for Meaning Distinction

1. There is (at least) one near-synonym for each sense of the adjective.
2. There is a set of context nouns that form grammatical constructions both with the original adjective and with the near-synonym.
3. The two sets of context nouns that characterize the different senses are non-overlapping.

Four Distributional Criteria for Meaning Distinction

1. There is (at least) one near-synonym for each sense of the adjective.
2. There is a set of context nouns that form grammatical constructions both with the original adjective and with the near-synonym.
3. The two sets of context nouns that characterize the different senses are non-overlapping.
4. The non-overlapping set of nouns forms a semantic category, reflecting the sub-selectional properties of adjectives (Pustejovsky, 1995).

Four Distributional Criteria for Meaning Distinction

1. There is (at least) one near-synonym for each sense of the adjective.
2. There is a set of context nouns that form grammatical constructions both with the original adjective and with the near-synonym.
3. The two sets of context nouns that characterize the different senses are non-overlapping.
4. The non-overlapping set of nouns forms a semantic category, reflecting the sub-selectional properties of adjectives (Pustejovsky, 1995).

Examples

- **Sense 1**: *napfényes* 'sunny', *napsütéses* 'sunshiny'
  Context nouns: *vasárnap* 'Sunday', *nap* 'day' ⇒ TIME
- **Sense 2**: *napfényes* 'sunny', *napsütötte* 'sunlit'
  Context nouns: *terület* 'area', *sziget* 'island', *oldal* 'side', *terasz* 'terrace'⇒ PLACE

Structure

1. Motivation
2. Problem: when to tell apart meanings in the case of polysemy?
   - Near-synonymy
3. Solution: distributional criteria for meaning distinction
4. Modelling meaning distinction
   - Representation of adjectives, adjectival meanings, polysemies
   - Detecting the subcategorized nouns
5. Method
6. Conclusion and future work

Modelling the Phenomenon: Adjectives

Representation of adjectives

Static word embeddings

Technical parameters:

- word2vec (CBOW)
- trained on cc. 170M sentences
- vector representations for cc. 8.5M wordforms
- window-size: 6
- min. frequency: 3
- Gensim python package

HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Modelling the Phenomenon: Adjectives

Representation of adjectives

Static word embeddings

Technical parameters:

- word2vec (CBOW)
- trained on cc. 170M sentences
- vector representations for cc. 8.5M wordforms
- window-size: 6
- min. frequency: 3
- Gensim python package

- Pros:
  - Easy to train and handle
- Cons:
  - Meaning Conflation Deficiency: "the inability to discriminate among different meanings of a word" (Camacho-Collados & Pilehvar, 2018)
  - $\Rightarrow$ A solution is needed

Modelling the Phenomenon: Adjectival Meanings

Solution for Meaning Conflation Deficiency

Static word embeddings $\Rightarrow$ graph:

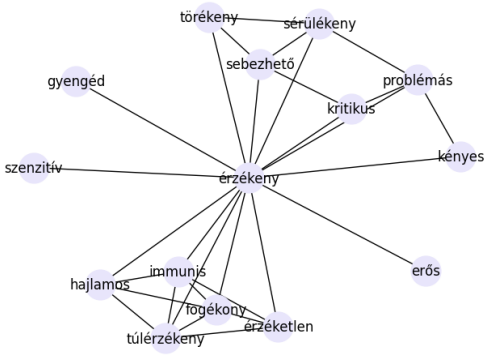
- adjectives $\rightarrow$ nodes
- semantic similarity $\rightarrow$ edges

Modelling the Phenomenon: Adjectival Meanings

Solution for Meaning Conflation Deficiency

Static word embeddings ⇒ graph:

- adjectives → nodes
- semantic similarity → edges



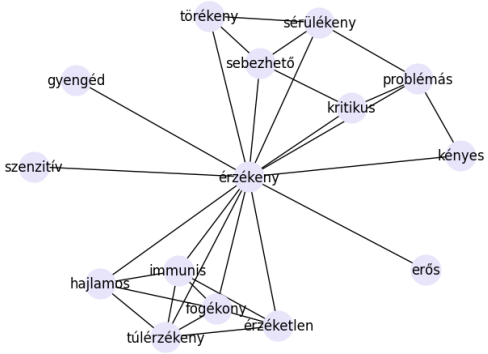
Modelling the Phenomenon: Adjectival Meanings

Solution for Meaning Conflation Deficiency

Static word embeddings ⇒ graph:

- adjectives → nodes
- semantic similarity → edges

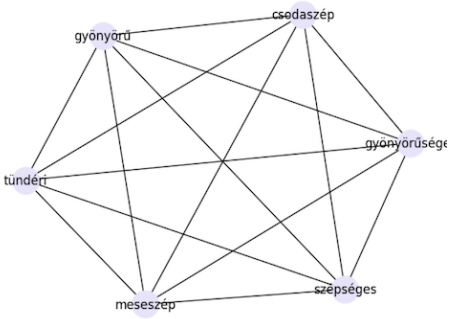
- Densely connected subgraphs indicate submeanings of *érzékeny* 'sensitive':
  - *hajlamos* 'prone to/tend to'
  - *fogékony* 'receptive'
  - *túlérzékeny* 'oversensitive'
  - *immunis* 'immune'
  - *érzéketlen* 'insensitive'
  - *sebezhető* 'susceptible'
  - *sérülékeny* 'vulnerable'
  - *törékeny* 'fragile'



Modelling the Phenomenon: Adjectival Meanings as Cliques

Densely connected (sub)graphs indicate (sub)meanings:

- Cliques: maximally connected:

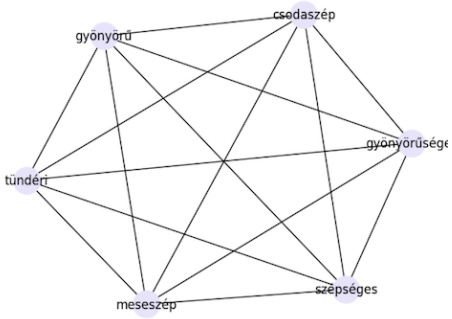


'beautiful', 'stunning', 'gorgeous', 'lovely', 'fabulous', 'adorable'

Modelling the Phenomenon: Adjectival Meanings as Cliques

Densely connected (sub)graphs indicate (sub)meanings:

- Cliques: maximally connected:
  - Every pair of nodes is connected

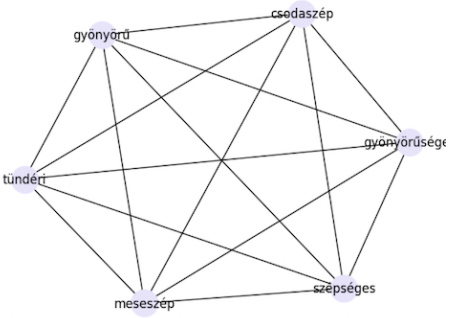


'beautiful', 'stunning', 'gorgeous', 'lovely', 'fabulous', 'adorable'

Modelling the Phenomenon: Adjectival Meanings as Cliques

Densely connected (sub)graphs indicate (sub)meanings:

- Cliques: maximally connected:
  - Every pair of nodes is connected
- The meaning of each element is similar to that of every other element



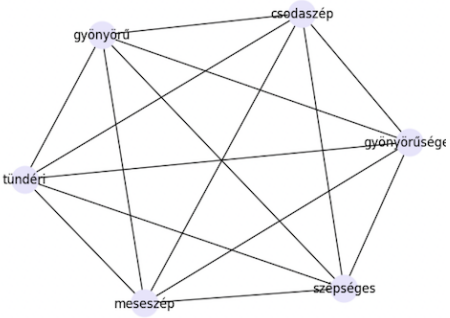
'beautiful', 'stunning', 'gorgeous', 'lovely', 'fabulous', 'adorable'

HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Modelling the Phenomenon: Adjectival Meanings as Cliques

Densely connected (sub)graphs indicate (sub)meanings:

- Cliques: maximally connected:
  - Every pair of nodes is connected
- The meaning of each element is similar to that of every other element
- ⇒ Cliques are modelling near-synonymy classes representing a (sub)sense of an adjective.



'beautiful', 'stunning', 'gorgeous', 'lovely', 'fabulous', 'adorable'

HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Modelling Adjectival Polysemy

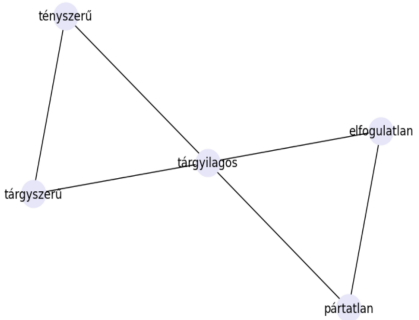
Adjectival polysemy

An adjective has multiple senses, if it belongs to multiple cliques (representing various near-synonymy classes).

Modelling Adjectival Polysemy

Adjectival polysemy

An adjective has multiple senses, if it belongs to multiple cliques (representing various near-synonymy classes).



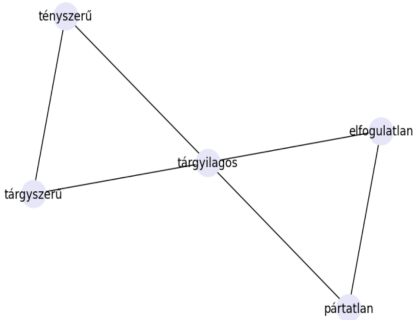
HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Modelling Adjectival Polysemy

Adjectival polysemy

An adjective has multiple senses, if it belongs to multiple cliques (representing various near-synonymy classes).

- *tárgyilagos* 'objective' →
  - Clique$_1$: {*tárgyszerű* 'concise'; *tényszerű* 'factual'}
  - Clique$_2$: {*pártatlan* 'impartial'; *elfogulatlan* 'unbiased'}



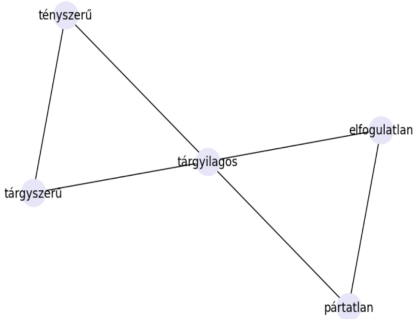
HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Modelling Adjectival Polysemy

Adjectival polysemy

An adjective has multiple senses, if it belongs to multiple cliques (representing various near-synonymy classes).

- *tárgyilagos* 'objective' →
  - Clique$_1$: {*tárgyszerű* 'concise'; *tényszerű* 'factual'}
  - Clique$_2$: {*pártatlan* 'impartial'; *elfogulatlan* 'unbiased'}
- Additional critera: Non-overlapping sets of nouns (2, 3, 4)



HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Clique Validation via the Following Nouns

(2) There is a set of context nouns that form grammatical constructions both with the original adjective and with the near-synonym.

(3) The two sets of context nouns that characterize the different senses are non-overlapping.

(4) The non-overlapping set of nouns forms a semantic category, reflecting the sub-selectional properties of adjectives (Pustejovsky, 1995).

Examples

- **Sense 1**: *napfényes* 'sunny', *napsütéses* 'sunshiny'
  Context nouns: *vasárnap* 'Sunday', *nap* 'day' ⇒ TIME
- **Sense 2**: *napfényes* 'sunny', *napsütötte* 'sunlit'
  Context nouns: *terület* 'area', *sziget* 'island', *oldal* 'side', *terasz* 'terrace' ⇒ PLACE

Detecting the Salient Nominal Contexts

How to identify nouns inducing the relevant meanings?

Examples

{mindennapi} 'common' ⇒ {hétköznapi} 'common', 'ordinary'
⇒ {mindennapos} 'everyday'.

COMMON, ORDINARY

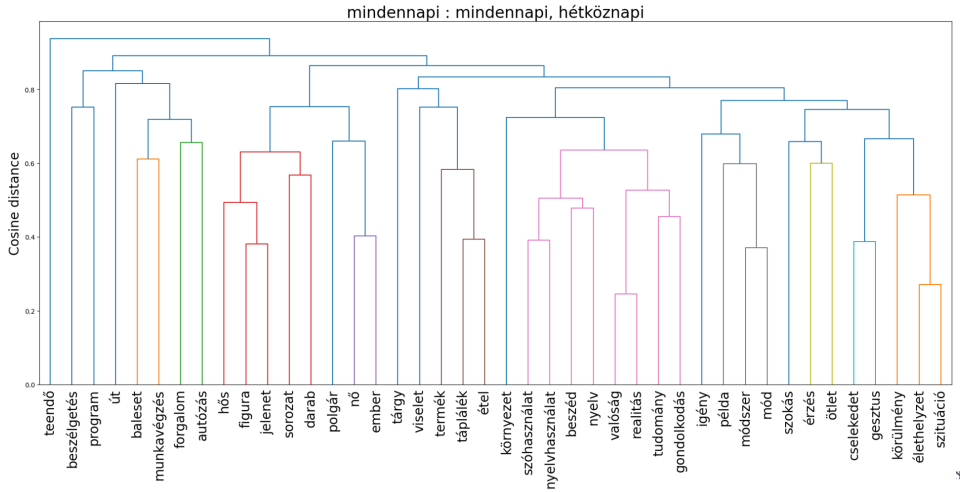
- *szóhasználat* 'word usage'
- *nyelvhasználat* 'language use'
- *valóság* 'reality'
- *tudomány* 'science'
- *gondolkodás* '(way of) thinking'

EVERYDAY

- *gyakorlás* 'practice'
- *testmozgás* 'exercise'

Detecting the Salient Nominal Contexts via Binary Trees — Dendrograms

How to identify nouns inducing the relevant meaning? — COMMON/ORDINARY



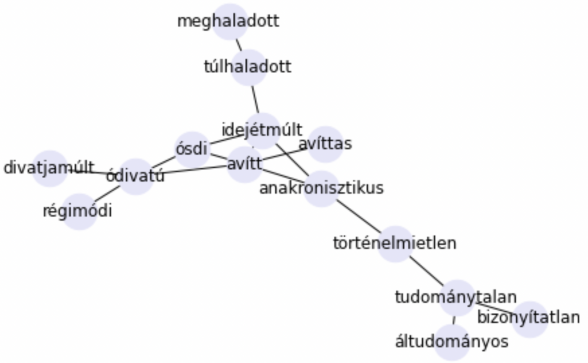
Semantic Domains as Connected Graph Components

- *Connected components* in the graph strictly corresponded to non-overlapping, semantically coherent components.
  - a **connected graph component** is a subset of network nodes such that there is a path from each node in the subset to any other node in the same subset
- The adjectival graph components
  - keep the various semantic domains separate
  - also reveal the relations between the inner node adjectives providing information on polysemies and meaning shifts
  - cliques emerge as parts of the connected components
- The original adjectival graph (10,153 adjs) was dissected into 1,807 components
  - a partition over 6,417 adjectives, where each component corresponds to a well-defined semantic domain.
  - one component of such networks is always a giant connected component, comprising approximately one-third of the input adjectives (3,736)

Semantic Domains as Connected Graph Components – An Example

Connected components offer lexicographers a neatly categorized headword list, enabling a more thesaurus-like editing process ($\Rightarrow\Leftarrow$ traditional alphabetical editing process)

- *idejétmúlt* 'outdated'
- *ósdi* 'shabby'
- *túlhaladott* 'obsolete'
- *anakronisztikus* 'anachronistic'.
- *történelmietlen* 'ahistorical',
- *áltudományos* 'pseudoscientific'



Structure

1. Motivation
2. Problem: when to tell apart meanings in the case of polysemy?
   - Near-synonymy
3. Solution: distributional criteria for meaning distinction
4. Modelling meaning distinction
   - Representation of adjectives, adjectival meanings, polysemies
   - Detecting the subcategorized nouns
5. Method
6. Conclusion and future work

Unsupervised Extraction of Representations from Corpus Data

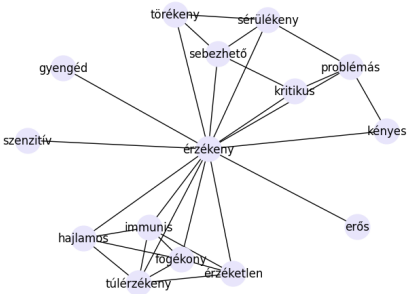
First step: generating weighted, undirected graphs

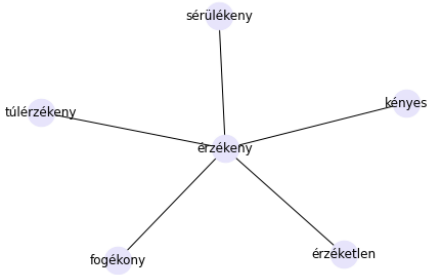
1. Graphs were created based on the word2vec representations of the adjectives
2. The nodes of the graph represent the adjectives
3. The weighted edges represent the semantic similarity between the nodes
4. The edge weights were calculated on the basis of the usual cosine similarity
5. The symmetric nature of cosine similarity guarantees that the adjectival graph is undirected

Unsupervised Extraction of Representations from Corpus Data

Second step: binarizing the weighted graph via a K cut-off parameter:

1. Edges are eliminated if $weight_i < K$
2. Edges are kept if $weight_i \geq K$
3. *Cliques* (polysemies) and *connected components* (semantic domains) were extracted from the resulting graph.



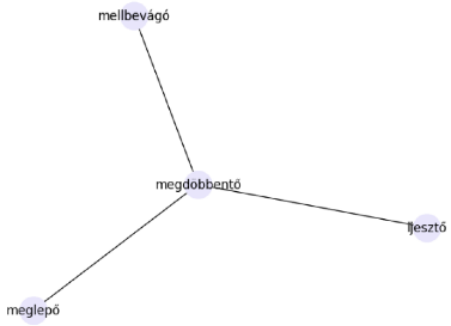


HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Testing Lexicographic Hypotheses – I

1. The induced cliques can help lexicographers to set up the adjectival microstructure.
   - The detailed analysis of the ego graphs of 20 frequent adjectives sliced at $K = 0.7$ showed that in 8 cases the corresponding cliques comprised relevant adjectives not in EDHL.

- Eg. the headword *megdöbbentő* lacks the subsense *mellbevágó*
  - *megdöbbentő* 'astonishing'
  - *ijesztő* 'frightening'
  - *meglepő* 'surprising'
  - *mellbevágó* 'gut-wrenching'



Testing Lexicographic Hypotheses – II

2. The automatically extracted nominal clusters, depicted by the dendrograms, provide lexicographers with additional contextual data to further characterize the exisiting adjectival microstructure in EDHL

   - It has been proved to be completely correct on the basis of randomly selected dendrograms.
   - This is due to the fact that nodes in the dendrogram near the terminals correspond to coherent, tight semantic classes of nouns
   - For instance, *fontos* 'important' may collocate MILITARY EVENTS (*csata* 'battle', *hadművelet* 'military operation', *küldetés* 'mission') or different types of ACTS IN LAW (*rendelet* 'order', *törvénytervezet* 'legislative proposal', *egyezmény* 'convention', *szerződés* 'contract')

Testing Lexicographic Hypotheses – III

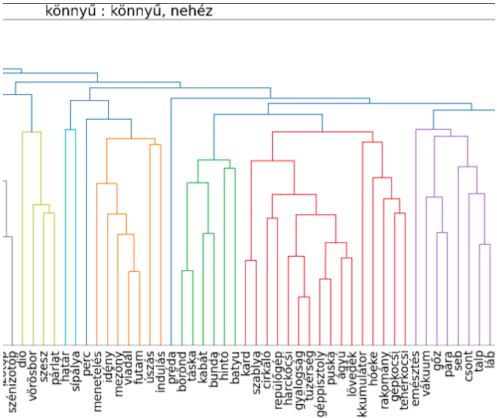
3. The nominal clusters characterizing the adjectival microstructure indicate on their own where meaning distinctions have to be made, without relying on any previously given definition

   - It has been proved to be only partially correct
   - Only dendrogram nodes near the terminals – in terms of cosine distance – indicate proper meaning distinctions
   - $\Rightarrow$ An additional filtering on validating nouns may be in order here

Example

Military-related LIGHT WEAPONs were differentiated in EDHL as follows:

- **Sense 1**: 'a <smaller-sized weapon> that does not require great effort to carry, transport, and handle'
- **Sense 2**: 'a <military unit> equipped with such weapons'



- *kard* 'sword',
- *szablya* 'saber',
- *puska* 'rifle',
- *ágyú* 'cannon',

- *gyalogság* 'infantry',
- *tüzérség* 'artillery'

Testing Lexicographic Hypotheses – IV

4. The automatically extracted connected components help to detect missing headwords, thus complementing the existing macrostructure
    - The comparison of the EDHL and the automatically retrieved, semantically related adjectives extracted via the connected graph components was rather conclusive
    - For instance, the graph-based algorithm cataloged 90 adjectives referring to quantities on the basis of the training corpus, out of which only 8 are listed in EDHL

Structure

1. Motivation
2. Problem: when to tell apart meanings in the case of polysemy?
   - Near-synonymy
3. Solution: distributional criteria for meaning distinction
4. Modelling meaning distinction
   - Representation of adjectives, adjectival meanings, polysemies
   - Detecting the subcategorized nouns
5. Method
6. Conclusion and future work

Conclusion and Future Work

- Unsupervised graph-based methodology to characterize both the adjectival macro- and microstructure in monolingual dictionaries
- The optimal value of the slicing parameter K should be set so that the automatically obtained results best suit the specific objectives of the lexicographers
- Nominal contexts should be also filtered $\Rightarrow$ an optimal frequency threshold should be set
- Scope of adjectives to be included in the dictionary
- Finally, the prototype algorithm should be implemented as a software tool to enhance the efficiency of lexicographers' work.

Thank you for your attention!