Determining the relationship between uninflectedness, overabundance and defectiveness

A contribution from distributional semantics

Dunstan Brown[1], Harald Baayen[2], Neil Bermel[3], Yu-Ying Chuang[2], Roger Evans[1], Alexandre Nikolaev[4]

[1] University of York
[2] Eberhard-Karls University of Tübingen
[3] University of Sheffield
[4] University of Eastern Finland

DGfS 2023, Workshop on Uninflectedness, March 8, 2023

With thanks to the AHRC, UK (grant AH/T002859/1
'Feast and Famine: Confronting Overabundance and Defectivity in Language')

Structure and Questions

- Nominal case and number inflection in Slavic (Czech and Russian) and Finnish
- Two kinds of uninflectedness (and what lies in-between)
  - lexemes that are not yet integrated into the inflectional system ('uninflected nouns' in Czech and Russian)
  - lexemes that are beginning to acquire some morphology
  - lexemes with an inflection that signals a change in POS (adverbial forms in Finnish)
- The relationship between uninflectedness, overabundance and, potentially, defectiveness
- How do uninflected items of the different types fit into the distributional semantic space?

Background – Distributional Semantics

- **Distributional semantics** (DS) is a quantitative language modelling approach in which words[1] are represented by high-dimensional (numeric) vectors, called word *vectors* or *embeddings,* based on their distributional relationships with other words in very large corpora.
- A simple embedding represents a word by frequency counts of the 1000 most frequent words that co-occur within 5 words in the corpus. More advanced embeddings such as **skip gram** construct predictive models of surrounding words and use the model parameters to represent the word.
- Some embeddings such as **fasttext** (Bojanowski et al., 2017) also use subword information, such as character n-grams, to enrich the vectors.
- With sufficiently large corpora, such representations turn out to be amenable to statistical manipulations such as semantic synonymy and antonymy detection, and compositional (noun+verb) semantics .

[1] We use 'word' to mean 'wordform' throughout, unless explicitly stated otherwise.

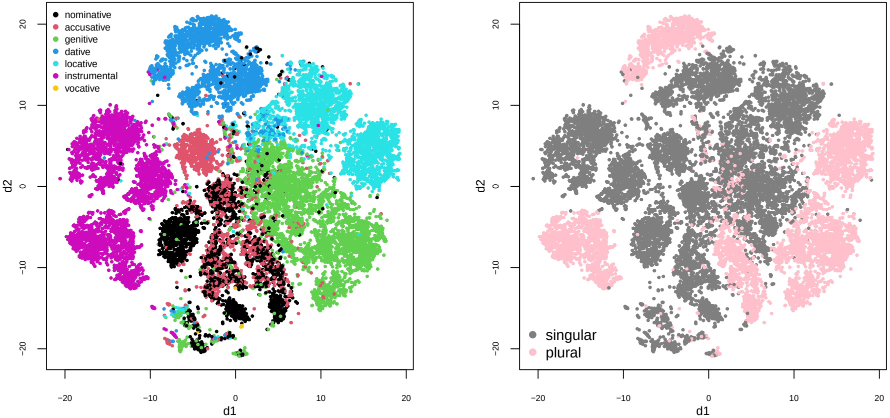
Background – Morphology and DS

- Although developed for semantics, DS is fundamentally holistic, making no explicit distinction between semantic, syntactic, morphological or any other aspects of word distribution.
- In particular, it is possible to explore morphological and morphosyntactic relationships from a distributional perspective, especially in languages with rich inflectional systems.
- In previous work, we found some evidence that defective nouns in Russian had distinctive distributional properties (Chuang et al., 2022)
- In this paper we explore whether similar techniques can be applied to overabundant and uninflected lexemes in Czech and Finnish.
- We use a range of statistical techniques, including **tSNE** dimension reduction (van der Maaten and Hinton, 2008)
- To facilitate detailed morphological exploration, we also work with corpora that are *morphologically disambiguated* (tagged)

Previous work – Defectiveness and DS

- Exploration of case and number distribution in Russian nouns using tSNE dimension reduction. Key conclusions:
    - tSNE clustering is sensitive to morphological token frequency – clusters into lexemes (micro-structure) when morphological range is limited, but morphological classes (macro-structure) eventually dominate as corpus size/variation increases
    - macro-structure shows case and number are not (distributionally) independent
- Distributional indicators for defectiveness (Russian noun genitive plurals):
    - small paradigm size
    - low semantic transparency
    - high error terms in decompositional models
- One limitation with this work is that it used raw wordforms rather than disambiguated forms

Case and Number Macro-Structure / Russian



Figures showing the (same) tSNE reduction of the 300D Russian noun embeddings to 2 dimensions. The left figure colours by *case*, showing clear clustering. The right figure colours by *number*. Within each *case* cluster, *number* also clusters, but not in a uniform pattern across all cases.

Evidence for Defectiveness

- Defectives tend to be low frequency
- Low frequency items in general tend to be more semantically transparent
- But defectives tend to be associated with lower semantic transparency, at least in Russian nouns that are defective (Chuang et al., 2022)

Uninflectedness – linguistic data / Russian

SG		PL	
NOM	*pal'to* (18.3%)	NOM	*pal'to* (5.5%)
ACC	*pal'to* (50.5%)	ACC	*pal'to* (4.1%)
GEN	*pal'to* (9.3%)	GEN	*pal'to* (3.1%)
DAT	*pal'to* (0.0%)	DAT	*pal'to* (0.0%)
PREP	*pal'to* (7.1%)	PREP	*pal'to* (0.5%)
INS	*pal'to* (1.6%)	INS	*pal'to* (0.0%)

The Russian noun *pal'to* 'coat'. Proportions of the total lexeme frequency (10,836) from the Araneum Russicum Russicum Maius Corpus (Benko, 2014), 859,319,823 words

Uninflectedness – linguistic data / Czech

SG		PL	
NOM	*rande* (˜20.2%)	NOM	*rande* (˜5.1%)
ACC	*rande* (˜45.5%)	ACC	*rande* (˜2.0%)
GEN	*rande* (˜3.0%)	GEN	*rande* (˜1.0%)
DAT	*rande* (˜1.0%)	DAT	*rande* (˜0.0%)
LOC	*rande* (˜20.2%)	LOC	*rande* (˜1.0%)
INS	*rande* (˜1.0%)	INS	*rande* (˜0.0%)

The Czech noun *rande* 'date'. Estimated proportions for uninflected realizations (based on sampling) of the total lexeme frequency (56,194) from the Czech csTenTen17 corpus (10.5 billion words) on Sketch Engine (Kilgarriff et al., 2014)

Czech 'growing' inflections

SG		PL	
NOM	*randeta* (0.6%), *randet* (0.4%)	NOM	*randata* (0.6%), *randeta* (0.4%)
ACC	*randeto* (0.2%)	ACC	*randata* (1.9%), *randeta* (0.6%)
GEN	*randete* (7.9%)	GEN	*randat* (0.9%), *randet* (0.9%)
DAT	*randeti* (3.2%)	DAT	*randetům* (0.4%)
LOC	*randeti* (40.7%)	LOC	*randetech* (1.1%), *randetích* (0.9%), *randetách* (0.2%)
INS	*randem* (31.9%), *randetem* (6.2%)	INS	*randety* (0.6%), *randaty* (0.2%)

The Czech noun *rande* 'date'. A further 469 instances of inflected forms associated with *rande*. From the Czech csTenTen17 corpus on Sketch Engine (Kilgarriff et al., 2014)

kotě - ’kitten’

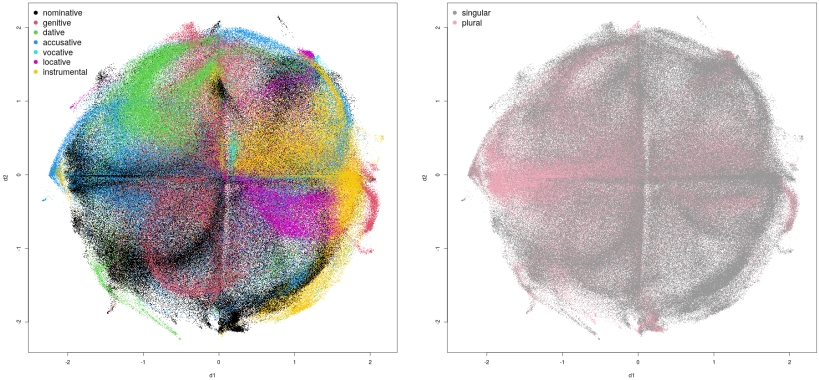
SG		PL	
NOM	*kotě*	NOM	*kot’ata*
ACC	*kotě*	ACC	*kot’ata*
GEN	*kotěte*	GEN	*kot’at*
DAT	*kotěti*	DAT	*kot’atům*
LOC	*kotěti*	LOC	*kot’atech*
INS	*kotětem*	INS	*kot’aty*

The Czech noun *kotě* ‘kitten’.This declension appears to provide a (partial) model for innovation of new inflections for *rande*.

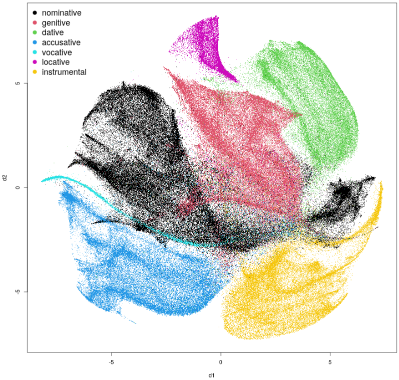
Uninflectedness → Overabundance

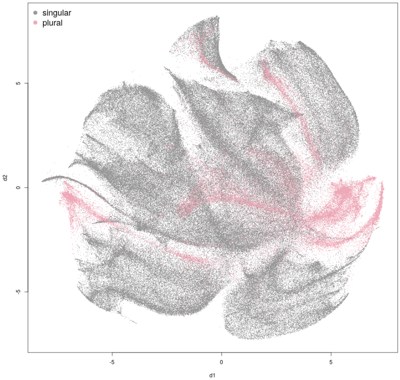
- Naranjo and Bonami (2021) identify two types of overabundance in Czech:
  - overabundance that is integrated into the inflectional system (locative singular)
  - overabundance that is orthogonal to the inflectional system (instrumental plural)
- Word embeddings for Czech (Kyjánek and Bonami, 2022) based on the SYN 9 corpus (Křen et al., 2021) (5.7 billion tokens)
- in Bermel and Nikolaev (2023) the difference between overabundance and defectivity is treated as a matter of uncertainty, the latter being considered a psychologically plausible way of considering defectiveness

Case and Number / Czech (all nouns)

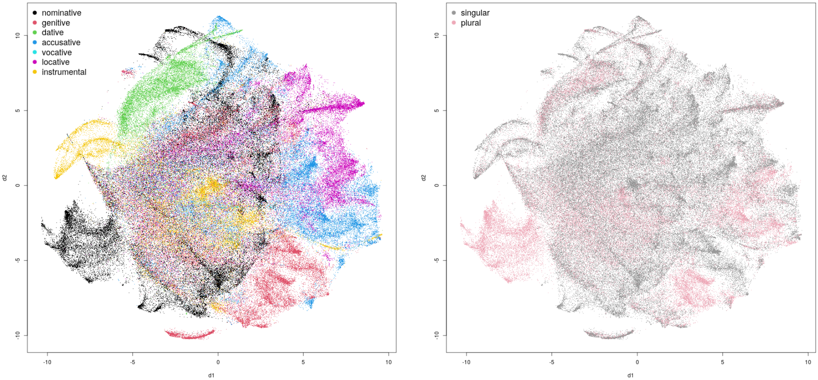


Case and Number / Czech (masculine – animate)

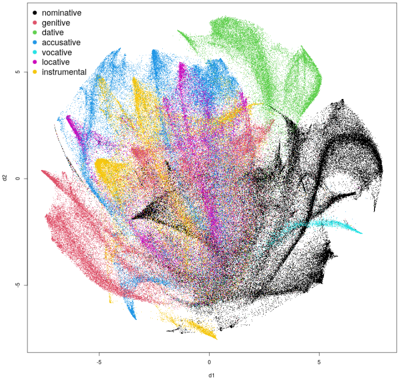


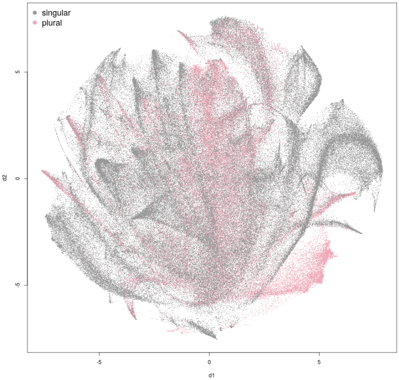


Case and Number / Czech (masculine – inanimate)

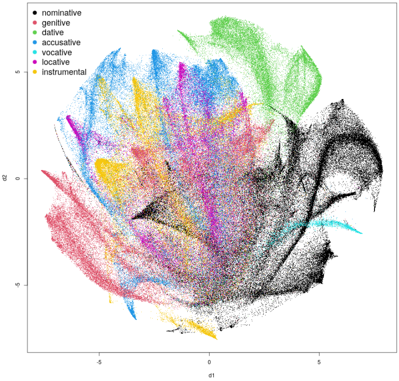


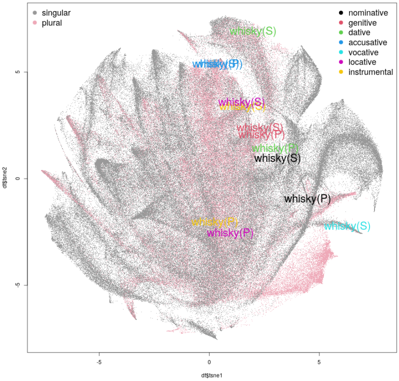
Case and Number / Czech (feminine)



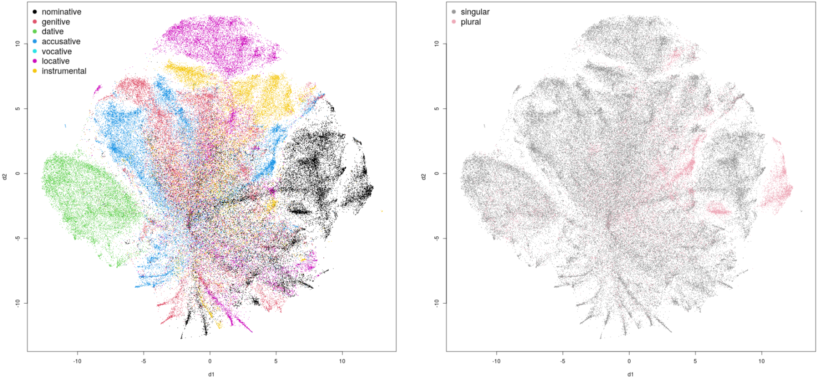


Case and Number / whisky (feminine)

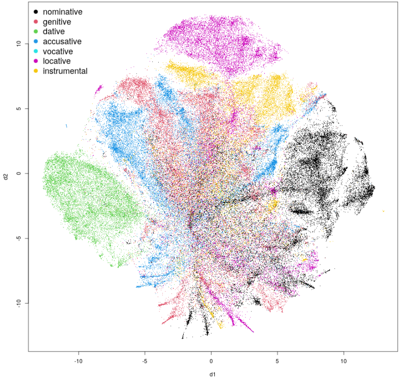


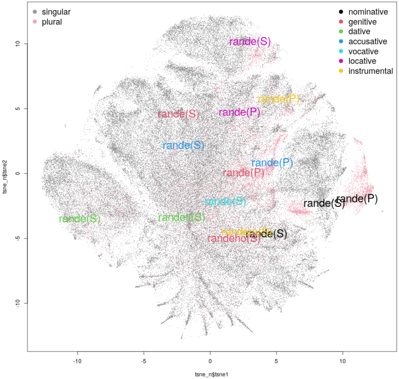


Case and Number / Czech (neuter)

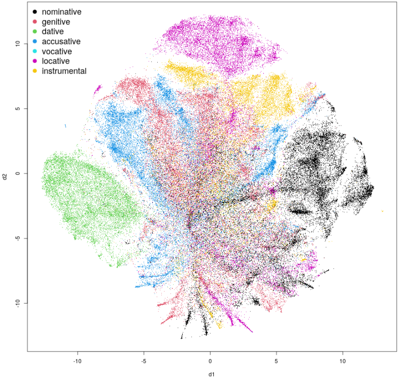


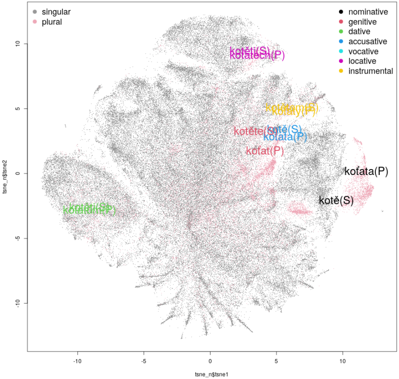
Case and Number / rande (neuter)



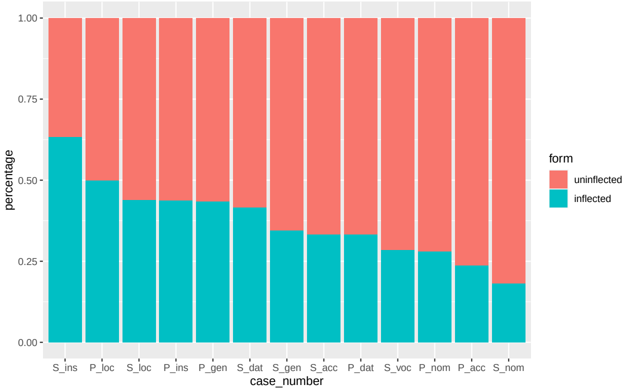


Case and Number / kotě (neuter)

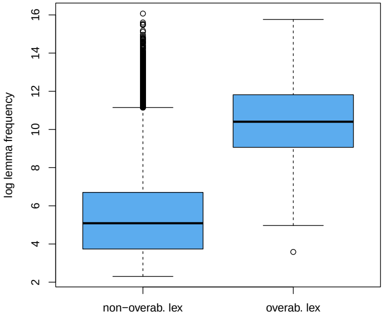




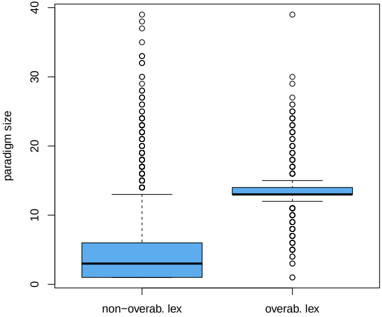
Form development



Lexemes with and without overabundant cells

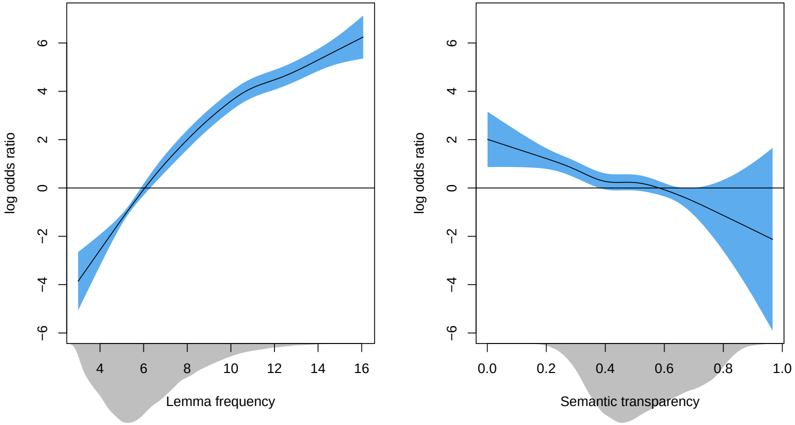


Lemma frequency

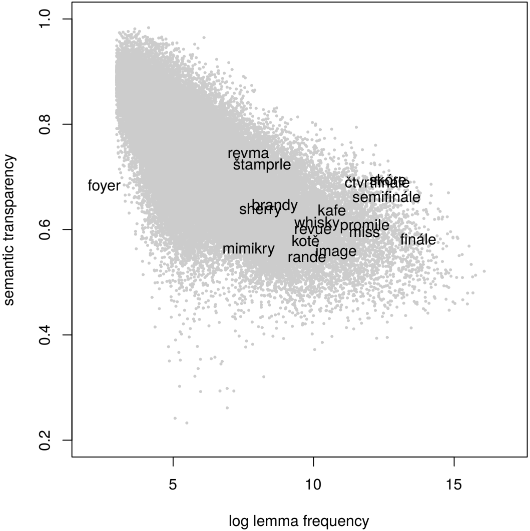


Paradigm size

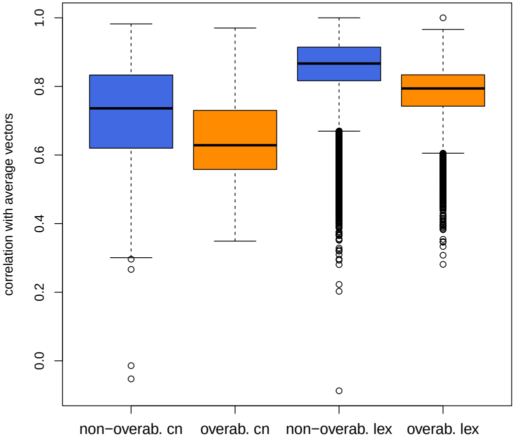
Predicting overabundant lexemes



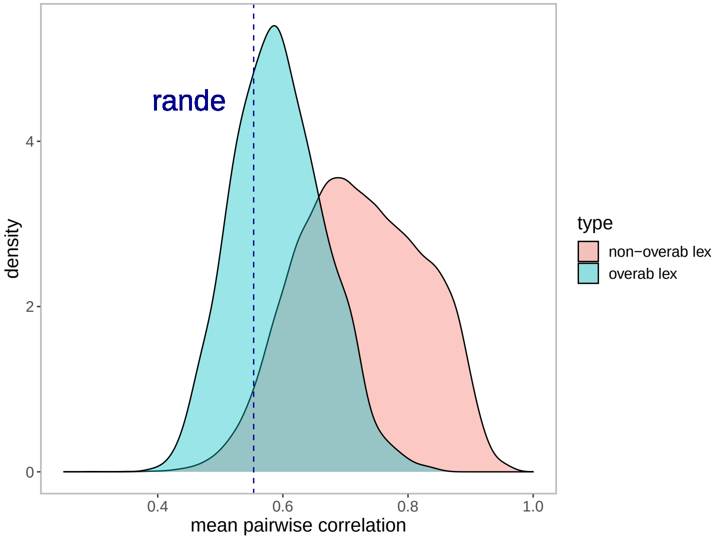
lemma frequency and semantic transparency



Correlations with average case_number and lexeme vectors



uninflectedness: rande



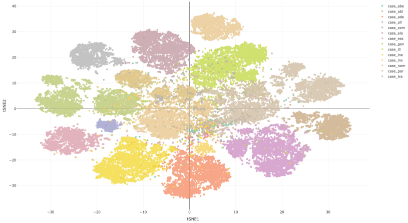
Interim Summary

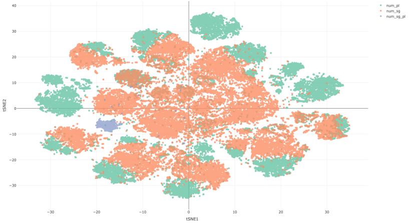
- Uninflected forms may start to 'grow' inflections
- When they do this, they become overabundant in some sense
- Overabundant items are less semantically transparent
- Perhaps for items such as *rande* they are not sufficiently frequent to avoid the uncertainty associated with some forms (e.g. gen pl)
- Other explanations are required to account for the form-based generalisations (particularly the adoption of the ins singular first)

Inflected to Uninflected?

- A brief look at Finnish pronouns and adverbial forms

Case and Number Macro-Structure / Finnish nouns





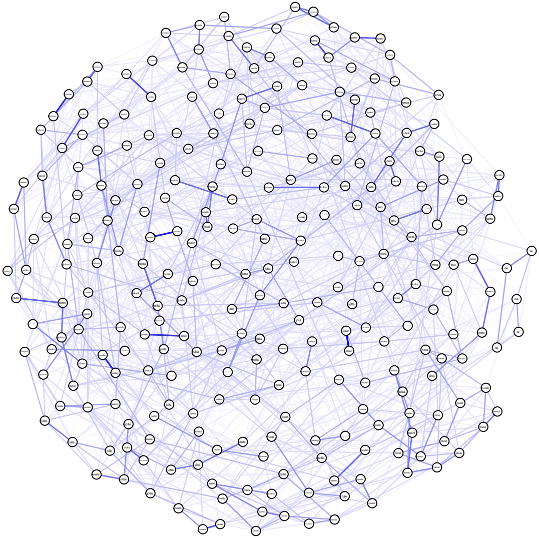
Finnish language

- In Finnish, inflected pronoun forms such as *tuolla* (*tuo* 'that' + ablative singular) may be used as uninflected adverbial forms ('over there').
- There are also other deictic / demonstrative / locative adverbs made from the same pronoun *tuo* 'that': *tuolta*, *tuolla*, *tuossa*, *tuosta*, *tuohon* that happen to be homonymous with the pronoun forms.
- However, there is one particular adverb *tuonne* 'over there' which is not homonymous with any inflected form of the pronoun *tuo*, because it has its unique ending *-nne*.

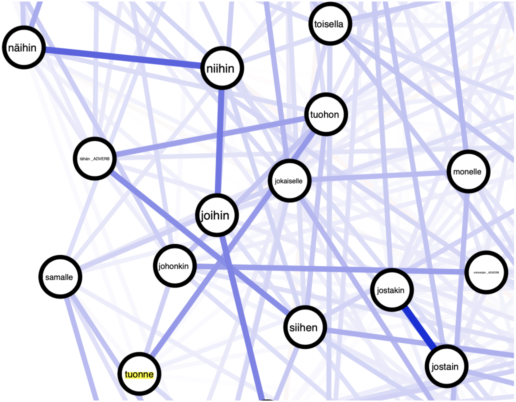
Finnish language

- The question arises: how well integrated will be semantic vectors (Fastext embeddings) of the pronoun *tuonne* in the network analyses of 370 most frequent pronoun forms?

Network analysis / Fasttext embeddings for 370 pronouns



Finnish language



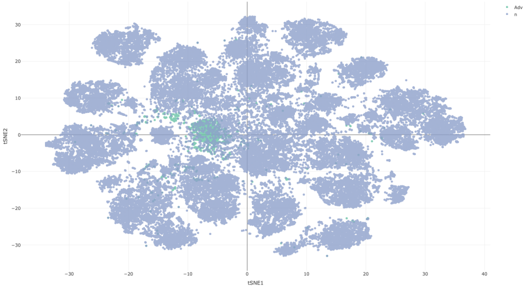
- Even though the adverb *tuonne* is located on periphery of the network, it seems to be well integrated with other pronouns by locative cases.

Finnish language

- From the perspective of the network based on Fasttext vectors, there is no clear-cut distinction between inflected pronouns and uninflected adverbs. They all behave as if adverbs are also inflected.
- Therefore, supposed uninflectedness of Finnish adverbs (e.g., corpus parsers do not label them with case) is not that obvious for the embeddings.
- Collocational profile of adverbs suggests that adverbs are structured in the embeddings according to case, or, in other words, case in some uses is not "dead" in the adverb; from a parser perspective, it does not make sense to assign case and number to adverbs, but then, all grammars leak (Sapir) and parsers cannot handle that very well; in other words, the semantics of "case" leak into the adverbs.

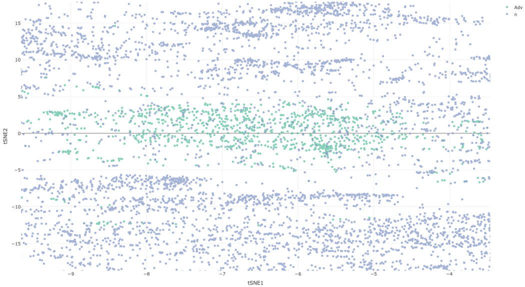
Finnish language

- When we combined 1420 adverbs with 55271 nouns in one tSNE plot, adverbs are clustered by a tSNE algorithm in their own relatively distinct cluster (green), which means that they are not clustered according to case clusters of nouns (blue clusters)



Finnish language

- However, adverbs (green dots) are mostly located between the two locative cases: between illative case words (upper blue dots, e.g., *asuntoon* 'into an apartment') and the inessive case words (bottom blue dots, e.g., *asunnossa* 'in an apartment').



Conclusion

- The patterns adopted by uninflected nouns when they start to 'grow' inflection are influenced by the available form sets (e.g. ins sg)
- Distributional methods help us understand what issues may arise when an uninflected moves to being inflected (via overabundance)
- Overabundant items are usually high frequent and low semantic transparency
- Could defectiveness be favoured where the overabundant item is insufficiently frequent? (e.g. gen pl of *rande*)
- Uninflected forms appear in the right location in the macro-structure, just as the Finnish adverbs appear close to their case-marked 'cousins'

References I

Benko, V. (2014). Compatible sketch grammars for comparable corpora. In Abel, A., Vettori, C., and Ralli, N., editors, *Proceedings of the 16th EURALEX International Congress*, pages 417–430, Bolzano, Italy. EURAC research.

Bermel, N. and Nikolaev, A. (2023). Uncertainty in the production of Czech noun forms.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chuang, Y.-Y., Brown, D., Baayen, R., and Evans, R. (2022). Paradigm gaps are associated with weird "distributional semantics" properties: Russian defective nouns and their case and number paradigm.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1:7–36.

References II

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kovářiková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., and Škrabal, M. (2021). SYN v9: large corpus of written czech. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kyjánek, L. and Bonami, O. (2022). Package of word embeddings of czech from a large corpus. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Naranjo, M. G. and Bonami, O. (2021). Overabundance and inflectional classification: Quantitative evidence from Czech. *Glossa: a journal of general linguistics*, 6(1).

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).