Community structure in inflectional networks

Maria Copot Andrea D. Sims

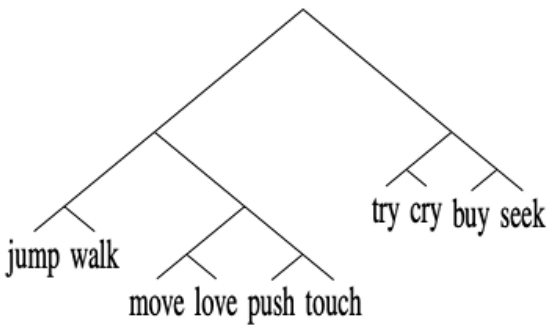
The Ohio State University

Inflectional classes

Inflectional classes are a linguistic **construct** used to describe **similarity in inflectional realisation** between lexemes

	MANGER *'eat'*	NAGER *'swim'*	FINIR *'end'*	PUNIR *'punish'*
1SG	mange	nage	finis	punis
2SG	manges	nages	finis	punis
3SG	mange	nage	finit	punit
1PL	mangeons	nageons	finissons	punissons
2PL	mangez	nagez	finissez	punissez
3PL	mangent	nagent	finissent	punissent

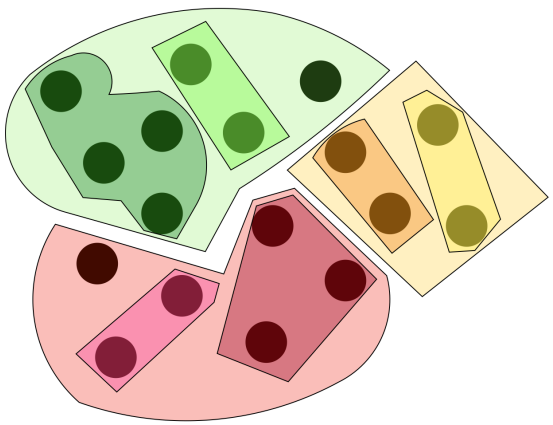
Describing inflectional systems



- Commonly observed that there is structure at multiple levels of specificity.

  Investigating relationship between structure at different levels through…
  - Clustering (Lee, 2014)
  - Connectionism, compression into hidden layers (Goldsmith & O'Brien, 2006)
  - Minimal description length (Beniamine et al., 2017)

What is the nature of the relationship between structure at different levels?



- Common **presupposition** that inflectional structure is fundamentally **hierarchical** (e.g. Dressler et al., 2006; Brown & Hippisley, 2012)
- Argued that hierarchy is **emergent from principles of learning and change** (Dressler et al., 2006; Stump, 2001)

- However, Beniamine (2021) shows that if one allows for multiple inheritance, **non-hierarchical structure appears common**.

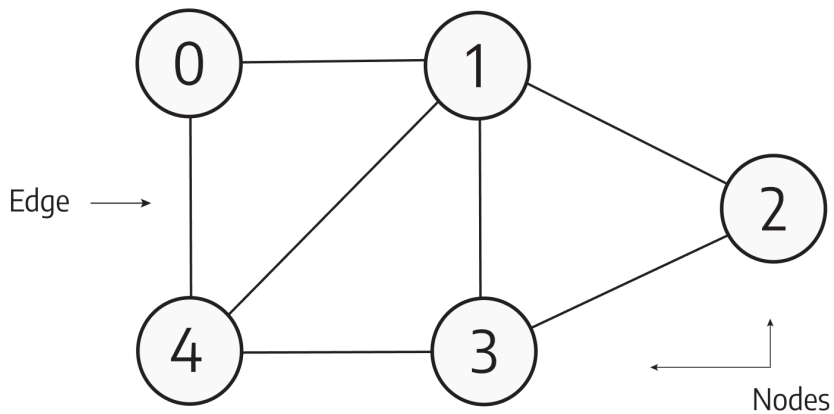
The thing that this is

- We investigate the hierarchy assumption in inflectional systems from a surface-oriented perspective
- We use network theory to
  - represent relationships of form in wordform paradigms
  - quantify the degree to which there is hierarchy in connectivity patterns
- We detect communities at different granularities
  - Are some partition sets more robust than others?
  - Are partitions organised hierarchically across levels?
  - Does the answer to these questions vary by language?

Network science

A network is a mathematical structure characterised by

- nodes (objects)
- edges (connections)



An approach that centers relationships between objects.

Network science

- Networks have been used to model system dynamics in the physical, biological and social sciences
- Limited adoption in modeling of lexical relations
  - e.g. Pham & Baayen (2015), Brown & Hippisley (2012)[1], Beniamine, (2021)[2], Sims (2020).

[1] Technically a tree
[2] Technically a semi-lattice

The linguistic systems

French verbs and Bosnian-Croatian-Montenegrin-Serbian (BCMS) nouns

A typical description of BCMS nominal inflection

	PROZOR ‘window (M)’	SELO ‘village (N)’	ŽENA ‘woman (F)’	NOĆ ‘night (F)’
NOM.SG	prozor	selo	žena	noć
GEN.SG	prozora	sela	žene	noći
DAT/LOC.SG	prozoru	selu	ženi	noći
ACC.SG	prozor	selo	ženu	noć
VOC.SG	prozore	selo	ženo	noći
INS.SG	prozorom	selom	ženom	noću
NOM.PL	prozore	sela	žene	noći
GEN.PL	prozora	sela	žena	noći
DAT/LOC.PL	prozorima	selima	ženama	noćima
ACC.PL	prozore	sela	žene	noći
VOC.PL	prozori	sela	žene	noći
INS.PL	prozorima	selima	ženama	noćima

But there are lots of “exceptions”

The four-class description abstracts away from a large amount of information about inflected nouns.

Excerpt from Šipka (2007):

b. The noun ends in –o/-e

If on the list of masculine –o/-e nouns, **then** apply that pattern

else if on the list of masculine Nom. –o/-e Gen.-e nouns, **then** apply that pattern,

else if on the list of masculine stem ending in –o, -e nouns, **then** apply that pattern,

else if on the list of masculine singular stem extension nouns, **then** apply that pattern,

else if on the list of neuter adjectival declension, **then** apply that pattern,

else if on the list of neuter interfix –et- nouns, **then** apply that pattern,

else if on the list of neuter interfix –en- nouns, **then** apply that pattern,

else if on the list of indeclinable feminine nouns ending in –e or -o, **then** do not decline the noun,

else apply default neuter pattern

A lossless partition of the system: 268 microclasses.

A typical description of French verbs

Three classes. Two defined by their infinitive and gerundive, a miscellaneous one for exceptions

	Class I	Class II	Class III	
	MANGER *'eat'*	PUNIR *'punish'*	SAVOIR *'to know'*	ALLER *'to go'*
INF	manger	punir	savoir	aller
GER	mangeant	punissant	sachant	allant
1SG	mange	punis	sais	vais
2SG	manges	punis	sais	vas
3SG	mange	punit	sait	va
1PL	mangeons	punissons	savons	allons
2PL	mangez	punissez	savez	allez
3PL	mangent	punissent	savent	vont

A lossless partition of the system (Beniamine, 2018): 97 microclasses.

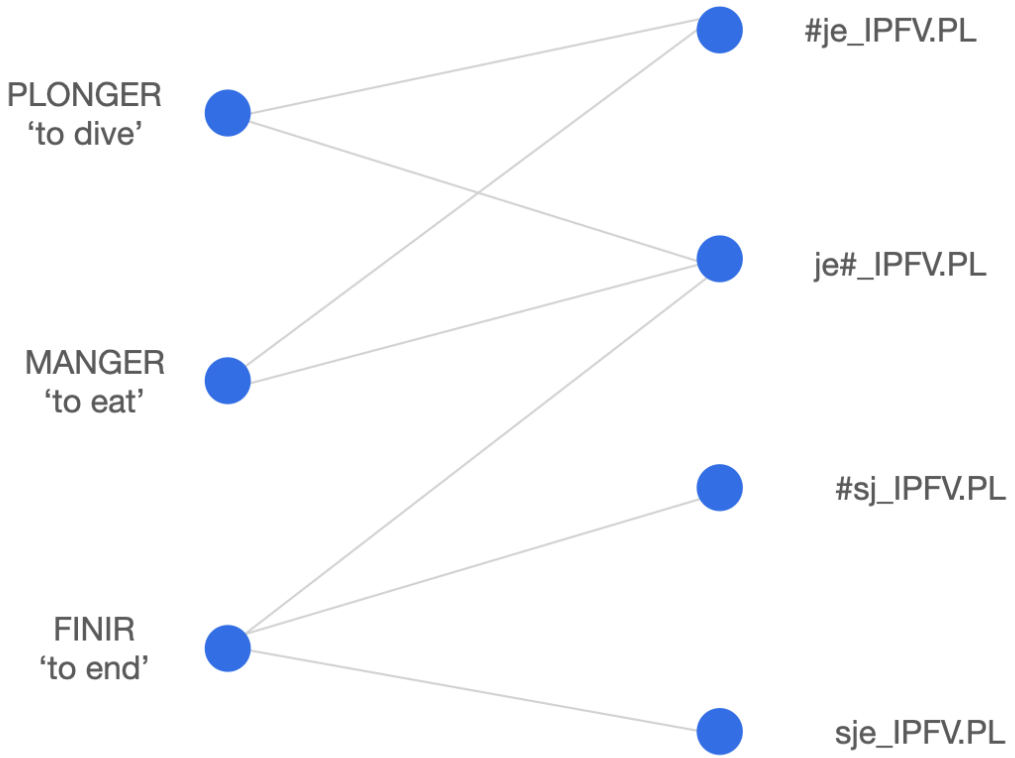
Setting up a network

Transposing an inflected lexicon to a network

1. Start with phonemically transcribed inflected lexicon
2. `morphalign` (Beniamine & Guzmán Naranjo, 2020) to align inflected forms.
3. `setmorph` (Carroll & Beniamine, submitted; Beniamine & Carroll, 2023) to segment inflected forms into smallest discriminative units (lexeme-internal comparisons)
4. Convert discriminative units into exponents:
    - combined exponents that were reliably adjacent
    - combine exponents for which a combined variant has been identified elsewhere in the system.
5. To capture subexponent regularities, the exponent string is segmented into triphones
6. The triphones are marked for the paradigmatic cell they occur in

Bipartite network

Two types of nodes: lexemes and exponent triphones.



	IPFV.2PL
PLONGER	plɔ̃ʒ-je
MANGER	mɑ̃ʒ-je
FINIR	fini-sje

Well-equipped to take into account gradient similarity between inflectional patterns.

The data

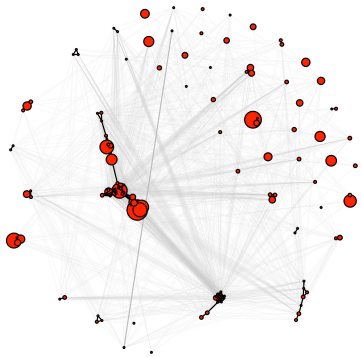
We create **bipartite** networks for

- **French verbs** (Vlexique 2.0, Beniamine et al. 2023)
- **BCMS nouns** (manually-corrected UniMorph Serbo-Croatian, Batsuren et al. 2022).

	French verbs	BCMS nouns
paradigm cells	52	12
unique lexemes	5.274	10.927
unique triphone + cell combinations	3.917	1.290
edges	605.896	202.999

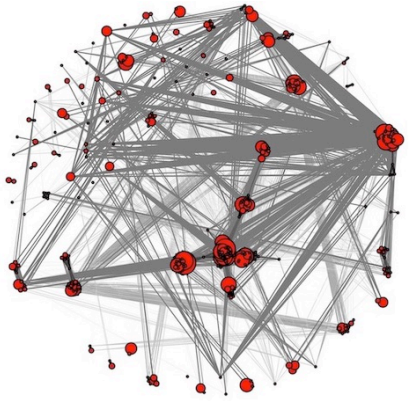
An International Space Station view of the systems

French



- A central cluster of microclasses.
- Most other classes are variations on the central cluster, characterised by the addition of unique exponents.

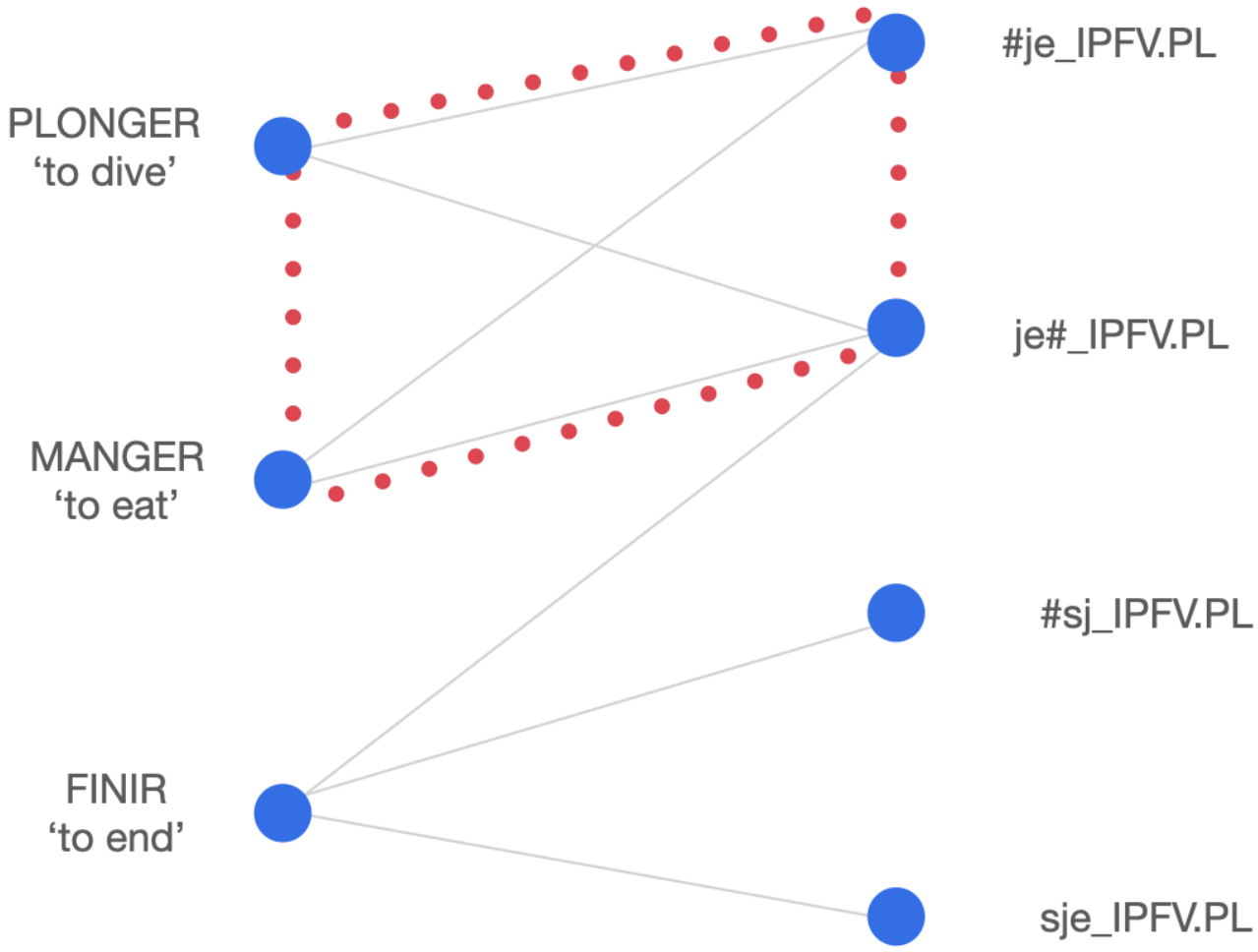
BCMS



- No clearly defined central cluster, high interconnectedness
- Quantifiably lower interpredictability of exponents (square clustering)

Square clustering

”The probability that my friends have common friends except me” - strength of joint probability of exponents.

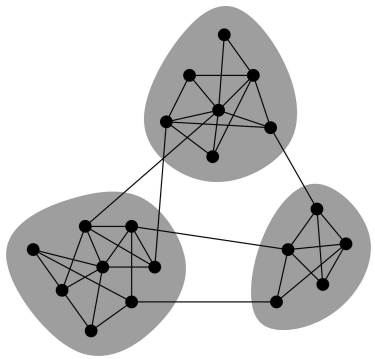


Inflection classes in networks

- An inflection class is a group of lexemes that share inflectional behaviour.
- In network terms, finding lexemes with shared inflectional behaviour is a community detection problem (Fortunato & Hric, 2016)
- Community detection is the process of grouping nodes in a network based on their internal connections, with the goal of finding clusters that are more densely connected internally than externally.
- Different structure exists at different levels of granularity, a parameter in community detection methods.

Community detection - Leiden algorithm

- We perform **community detection** over the bipartite network with the **Leiden algorithm** (Traag et al., 2019).
- The algorithm seeks to maximise **modularity**
  - find subsets of the graph where nodes are more connected with each other than with other nodes



- Method is known to yield communities with high **stability**.
- The **resolution parameter**
  - given time t, a **random walker** reliably does not escape the community.
  - Lower resolution = bigger communities.

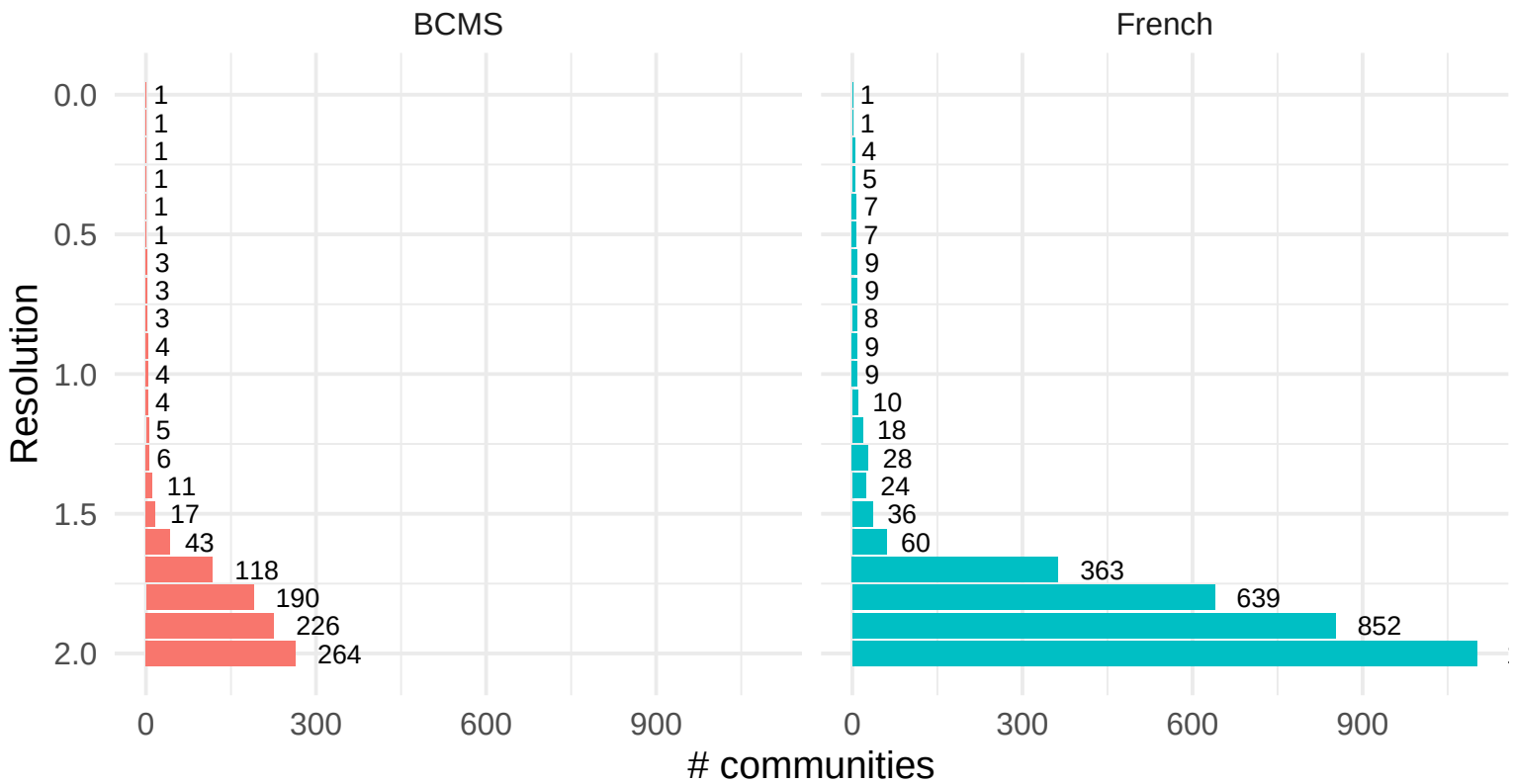
Community detection and inflection classes

- The resulting communities are conceptually comparable to inflection classes but not identical to them
- The same feature of inflectional similarity may have differing importance at different granularities
- Intuitively, partitions are built around exponents that serve as good discriminators for group membership

Finding community structure at different levels

- We set out to find community structure at different levels of granularity
  - number of communities is insensitive to granularity if strong IC organisation is an inherent property of the system
- We test the existence of hierarchical structure
  - Are communities at adjacent resolutions in super-/sub-set relationships, or are different factors important at different resolutions?
- While there may be multiple valid descriptions of a system, can we diagnose invalid descriptions by appealing to measure incoherence?

Number of communities

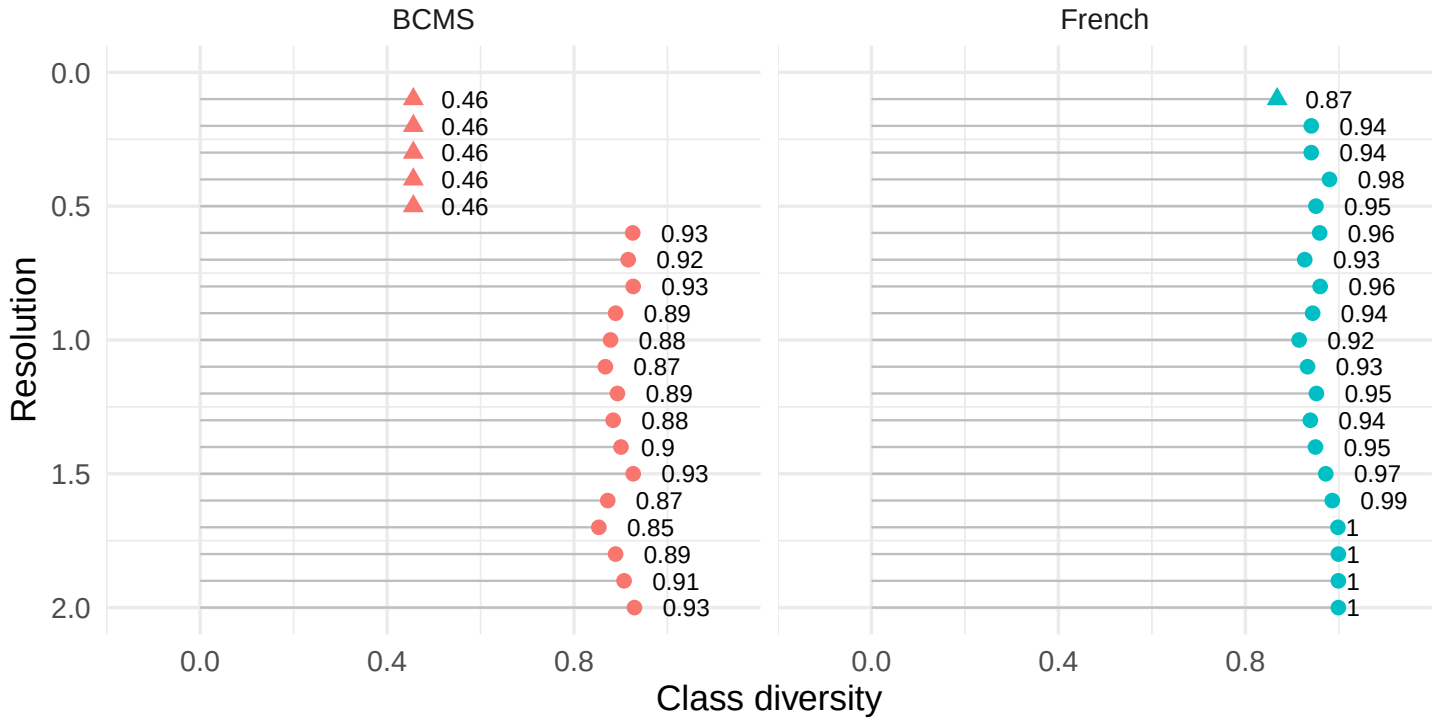


- BCMS hard to find partitions at first, slowly increasing in number
- French communities balloon for res > 1.6, very few lexemes per community.

Class diversity

- Is the method finding things that are crazy?
- We compare the partitions to the coarsest grouping, traditional classes.
- Class diversity score:
  1. For each partition, find traditional class most represented
  2. Calculate percentage of items of said class in the partition
  3. Average over partitions

Class diversity



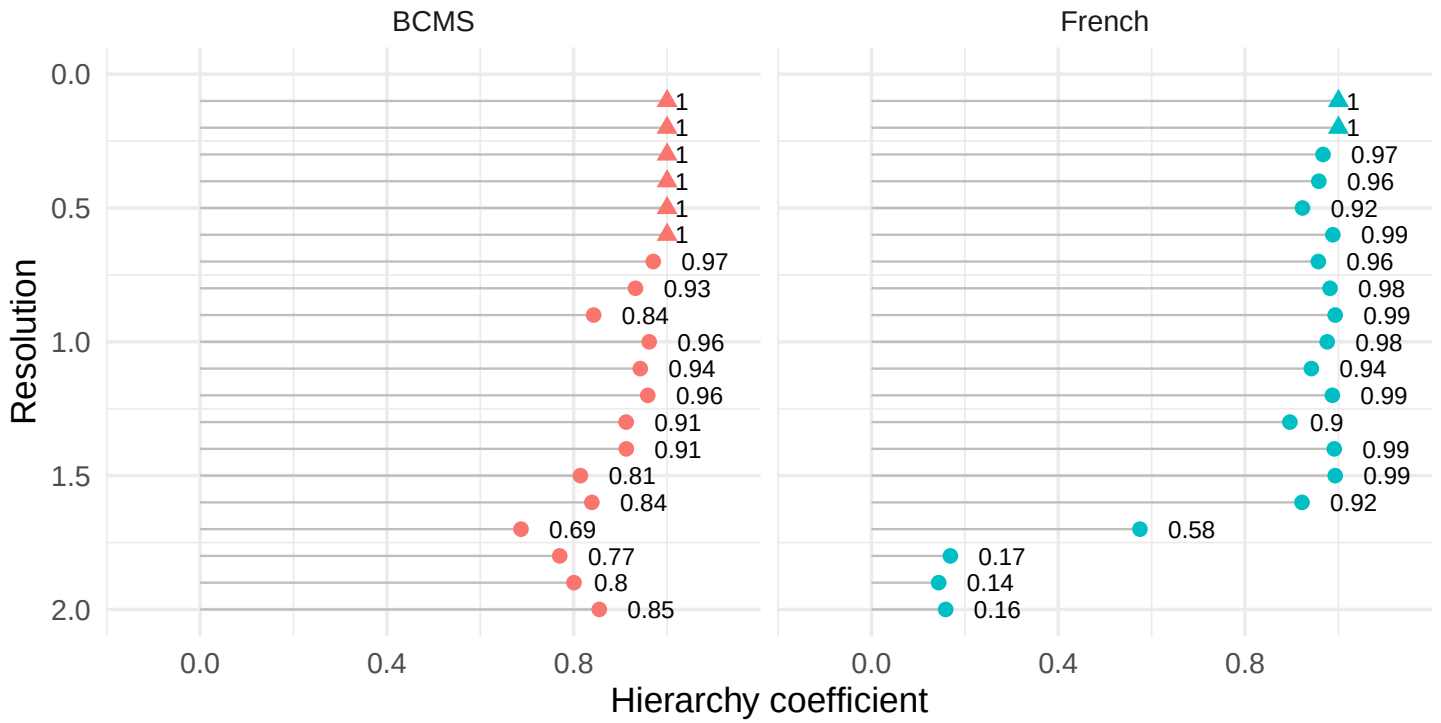
- Triangles are instances of only one community.
- 90% of lexemes in a given community belong to the same traditional class. BCMS < French.
- Score of 1 for French reflective of monolexemic communities

Hierarchy

- Are partitions at different granularities subsets/supersets of each other?
- Hierarchy score
  - For each pair of adjacent resolutions...
  - For each partition in the finer resolution...
  - For each pair of lexemes...

  1. Check what percentage of lexeme pairs in the finer resolution partition are in the same coarser resolution partition
  2. Average across lexeme pairs and partitions

Hierarchy



- Hierarchy holds somewhat. Nonmonotonic pattern.
- Sudden increases/drops for measures suggests that algorithm has difficulty finding generalisations.

Comparison with other approaches

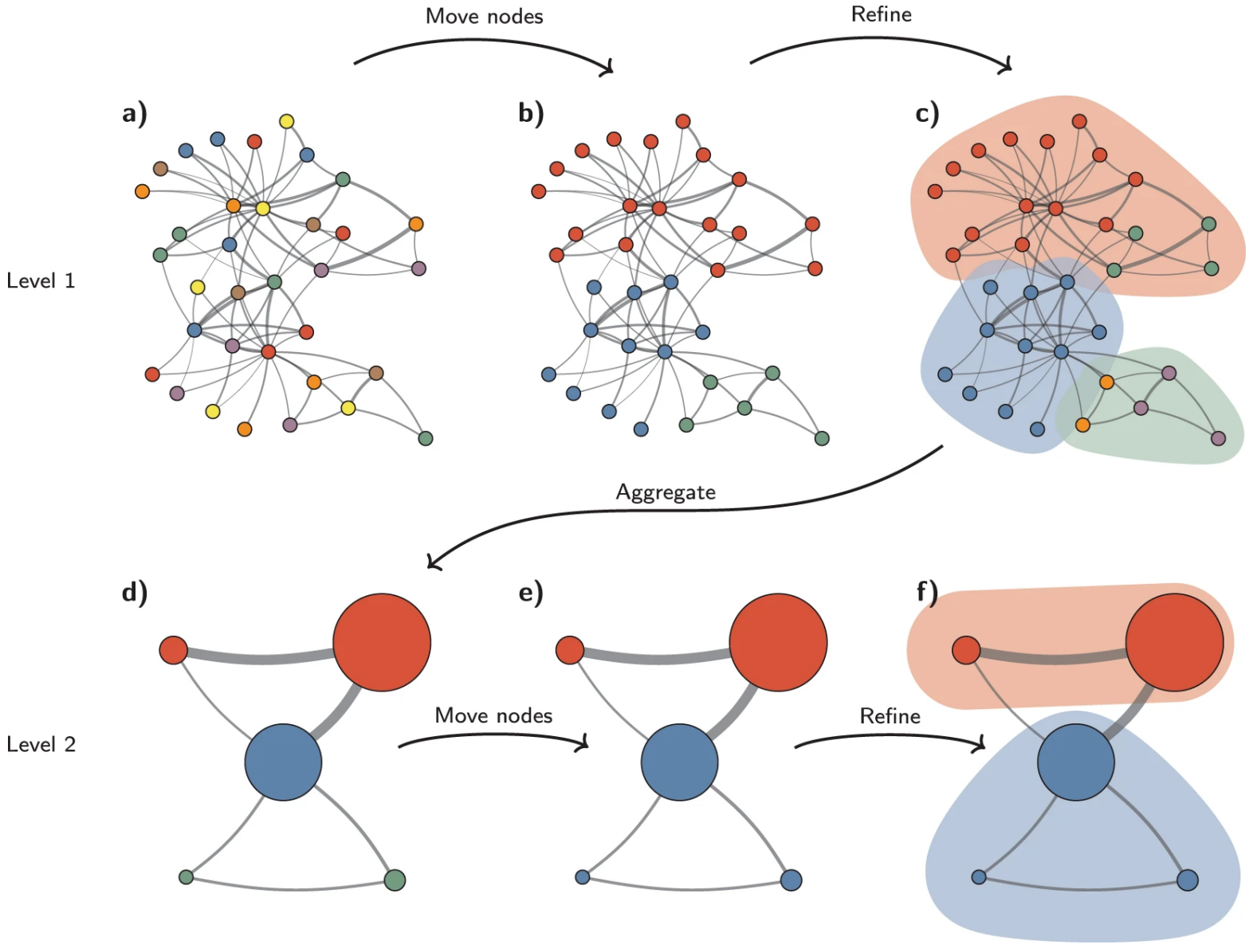
- The work follows recent interest in surface-oriented generalisations within WP approaches
- The hierarchical structure of inflectional systems is directly investigated, rather than assumed as a possiblity/necessity as part of the architecture (cf Dressler et al. 2006; Brown & Hippisley, 2012; Beniamine, 2021)
- A tool to compare the organisation of morphological systems

Conclusion

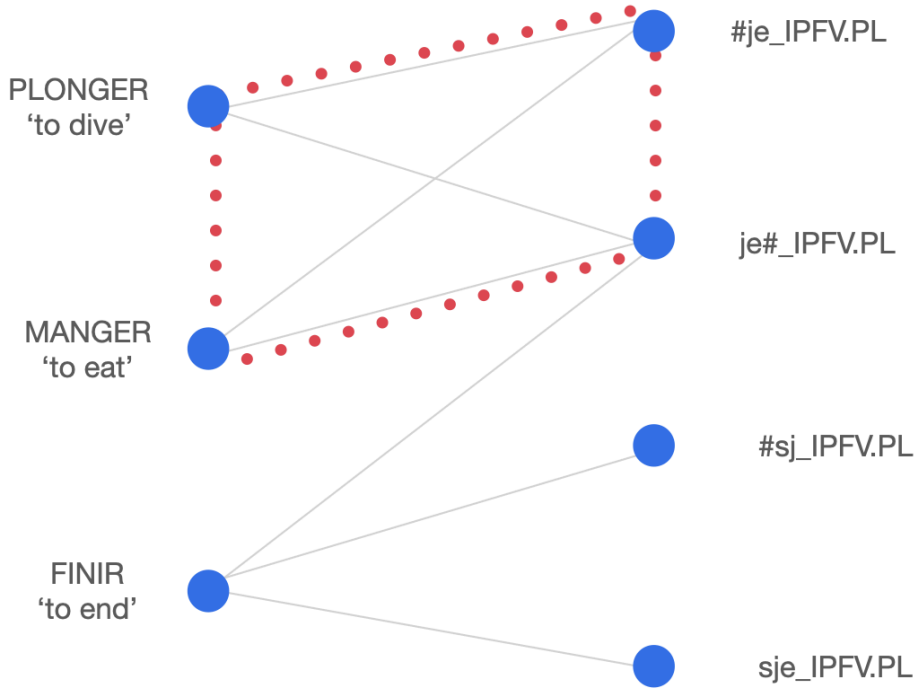
- A novel method for investigating inflectional structure through partitioning at different granularities
- The approach shows that stable structure exists at multiple levels of granularity
  - Multiple descriptions of a system are valid, and may be appropriate for different purposes
  - At too high granularity, method becomes unstable - certain descriptions are inappropriate
- While some hierarchy exists between resolutions, it is not absolute, and behaves nonmonotonically.

Appendix

Leiden algorithm



Square clustering

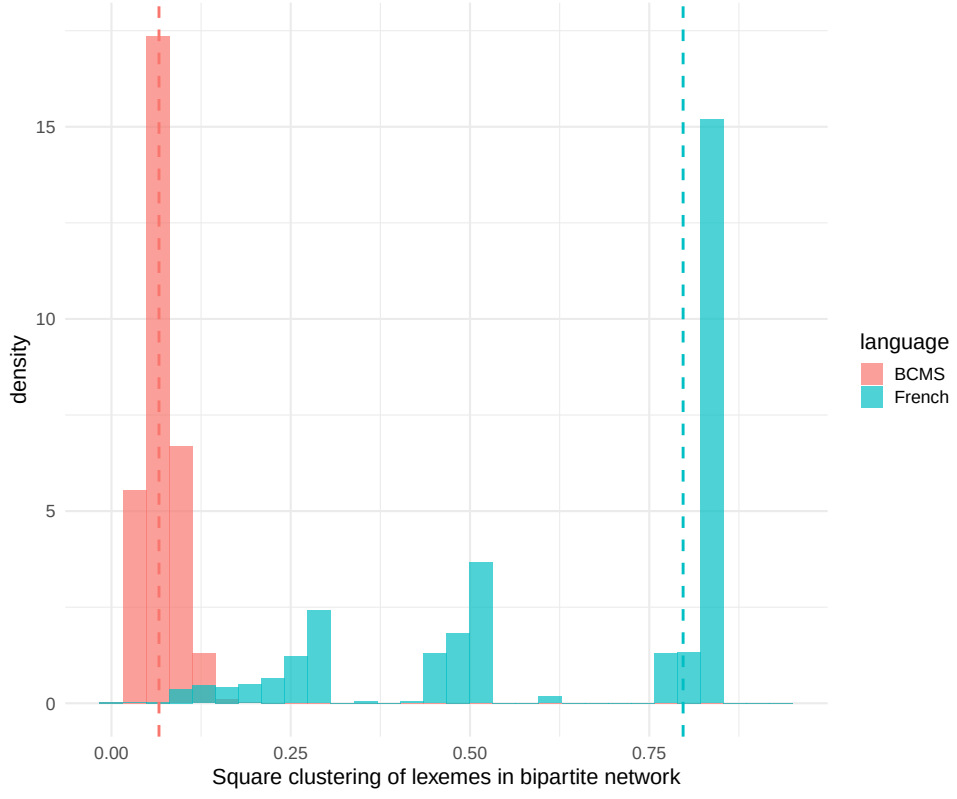


The network has one square. The SC coefficient measures how many squares a node is part of, out of the total possible number. In this network:

PLONGER	1/2 = 0.5
MANGER	1/2 = 0.5
FINIR	0

Square clustering in French verbs and BCMS nouns

The two systems have very different configurations: in French, exponents have much higher joint probability than in BCMS.



In French, on average, exponents are better at cueing the rest of a lexeme's paradigm.