Contrastive Morphological Typology and Logical Hierarchies

John Sylak-Glassman and Ryan Cotterell

Johns Hopkins University
Center for Language and Speech Processing

Friday, April 22, 2016
52nd Annual Meeting of the
Chicago Linguistic Society

Introduction

- A central goal of research in morphological typology:
  - Discover a list of categories (e.g. Tense) and features (e.g. present) that captures the concepts encoded by inflectional morphology across the world's languages.
- List would be a substantive claim about the content of the inflectional morphological component of grammar.
  - Would clarify interfaces of morphology with syntax and semantics.
- Need a principled way to discover these features and their organization.

Introduction

- *Claims*:
  1. Method based on overt contrast can be used to discover a set of semantically 'basic' inflectional features.
  2. These features are organized into logical hierarchies based on generality of meaning: Specific features are dominated by general ones.

1. Contrastive method of determining features
   - Definition of features and theoretical status
   - Summary of current feature set
2. Hierarchical organization of features: Logical hierarchies
   - Definition and sources of evidence
   - Grammaticalization paths
   - Cross-lingual morphological contrast mismatches

Contrastive Morphological Features: Method

- *Goal*: Find set of features that captures concepts encoded by inflectional morphology across the world's languages.
- *Guiding Idea*: Find most basic, "atomic" inflectional morphological distinctions that are never decomposed further in the world's languages.
  - More complex distinctions can be built additively or disjunctively.
- *Method*: If a morphosemantic distinction is encoded by two overtly contrasting morphemes in a language (of which one may be phonologically null) and the meaning encoded by at least one of the morphemes is not decomposed further in any other language, then that non-decomposable meaning is represented by a feature.
  - Like contrastive phonological features (Jakobson et al. 1952), but always privative and less abstract: e.g. SG (not [±MINIMAL])
- *Example*: Case marking for core arguments
  - English (pronouns): Two features, NOMINATIVE & OBLIQUE.
  - Later, observe languages which overtly mark following cases: ACCUSATIVE, ERGATIVE, ABSOLUTIVE, DATIVE, which cover functions of OBLIQUE → OBLIQUE not a feature.

Builiding a Contrastive Morphological Feature Set

- *Categories*:
  - Started by finding categories to which agreement features belong, then finding categories by primary part of speech.
- *Features*:
  - Examined languages with largest known number of distinctions (e.g. number in Sursurunga; Corbett 2000:26-30).
  - Found finest level of division within scalar distinctions (e.g. SG, DU, TRI, PAUC, GPAUC, PL) as well as other irreducible basic features (e.g. Arabic greater plural; ibid.:32).
- *Scope*: Limited to overt, affixal morphology and contrasts expressed paradigmatically.

Contrastive Morphological Features

- Features are intermediate between:
  - *Universal Category*: Universally available for any language, possibly psychologically 'real', used for description, analysis, and comparison
  - *Comparative Concept*: Defined by typologists, expressly for comparison, but cross-linguistically valid (Haspelmath 2010).
- *Contrastive Morphological Features*: Assumed to be universally available, have cross-linguistically consistent meaning to ensure comparability, not assumed to be psychologically 'real'
- Majority of features represent finest distinctions in meaning possible, cross-linguistically common groupings of basic features, e.g. subjunctive and irrealis, are also represented.
  - Unclear if these features have aspects that are basic, or if full back-off to "atomic" features would be equally expressive.

A Contrastive Morphological Feature Set

- Broad survey of typological literature resulted in set of over 277 features distributed among 25 morphological categories.
  - *Categories*: Aktionsart, animacy, argument marking, aspect, case, comparison/grade, definiteness, deixis, evidentiality, finiteness, gender, information structure, interrogativity, mood, number, POS, person, polarity, politeness, possession, switch-reference, tense, valency, voice
- Reasons for large feature count:
  - *Composite features*: Features for possession marking (27) are of form: *possession* + *person* + *number* + { *gender* / *clusivity* }
  - Diversity of organization of gender / noun-class systems: Many Bantu, Nakh-Daghestanian noun classes
- Used feature set (UniMorph Schema): Principled, accurate glossing for inflectional morphemes across languages.
  - Used in creating database of inflected word forms from all languages on English edition of Wiktionary. (← Used later to supply morph. features)
  - Full details in: Sylak-Glassman et al. (2015a,b) as well as *UniMorph Schema User Guide* (contact for copy).

Morphological Feature Organization

- Category Membership
- Hierarchical Organization: 1. Dependency Hierarchies
  - Existence of feature lower in the hierarchy entails existence of the feature which dominates it.
- Morphological Feature Geometry of Harley and Ritter (2002) encodes dependency through representational complexity: More complex representations entail simpler ones, but not vice versa.
- Similarities to phonological feature geometry (Clements 1985; Sagey 1990):
  - 'Substance' constraining morphological feature geometry is "conceptual in nature," and includes "notions such as deixis, countability, and taxonomy" (Harley and Ritter 2002:484-5).

Morphological Feature Organization

- Hierarchical Organization: 2. *Logical Hierarchies*
- Features lower in hierarchy specify the parameter encoded by the dominating feature in a more specific way.
- 'Substance' which constrains logical hierarchy is specificity-generality relationships which arise from content of features.
- Logical hierarchies can resolve cross-lingual morphological mismatches, explain facultative feature use and superclassing (Corbett 2012:21-26), and yield predictions about the results of the genesis or loss of morphological contrast.
  - In this talk, concentrating on logical hierarchies within a morphological category (e.g. number, case).
- Part II Roadmap:
  - Evidence overview
  - Grammaticalization
  - Cross-lingual morphological contrast mismatches
  - Predictions and Discussion

Evidence for Logical Hierarchies: Overview

1. *Language-internal*: Facultative feature use (Corbett 2012:21-22)
   - Larike-Wakasihu has SG, DU, TRI, PL, but speakers may use PL in place of DU or TRI
2. *Language-internal*: Superclassing (Corbett 2012:22-24)
   - Jingulu has four semantically-determined genders: Masculine, feminine, vegetable, neuter. Full agreement between N and ADJ is possible, but agreement can be reduced to using a masculine and neuter form, in which MASC = MASC + FEM and NEUT = NEUT + VEG. More abstractly, this is ANIM vs. INAN.
3. *Diachronic*: Grammaticalization pathways in which a generality relationship exists.
4. *Cross-lingual*: Among aligned words in parallel text from two languages, if one or more features in the higher contrast language maps to a feature in the lower contrast language, the feature in the lower contrast language likely dominates the feature(s) in the higher contrast language in a logical hierarchy.

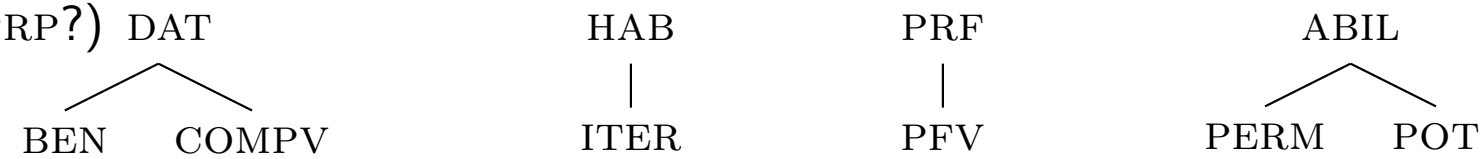
Grammaticalization: Survey Criteria

- Surveyed attested grammaticalization processes (source $\rightarrow$ target pairs) in Heine and Kuteva (2002), a database of 400+ processes.
- Limited survey to processes in which both source and target clearly belonged to morphological categories identified in UniMorph Schema. Further limited to processes in which both sides belonged to same category.
- Excluded processes in which one side was a free lexeme (e.g. 'say' $\rightarrow$ Hearsay Evidential) or purely syntactic category (e.g. VP-and $\rightarrow$ subordinator)
- Of 30 processes, 4 explicitly reported as involving a specificity-generality relationship, 12 more judged to by us.

Grammaticalization: Results

Reported S→G[1]	Judged S→G	Judged G→S
benefactive → dative	reflexive → anticausative	ability → permissive
benefactive → purpose	reflexive → reciprocal	ability → possibility
instrument → manner		ablative → material
iterative → habitual		ablative → partitive
		allative → until
		comitative → temporal
		conditional → concessive
		dative → comparative
		dative → patient
		perfect → perfective

- Arrow indicates diachronic directionality.
- 9 processes involve Case, 2 Valency, 3 Mood, 2 Aspect
- Evidence for following pieces of logical hierarchies:



- *Issues*: Should there be features for: Manner, anticausative, material, concessive, and patient? Need further evidence of contrasts.

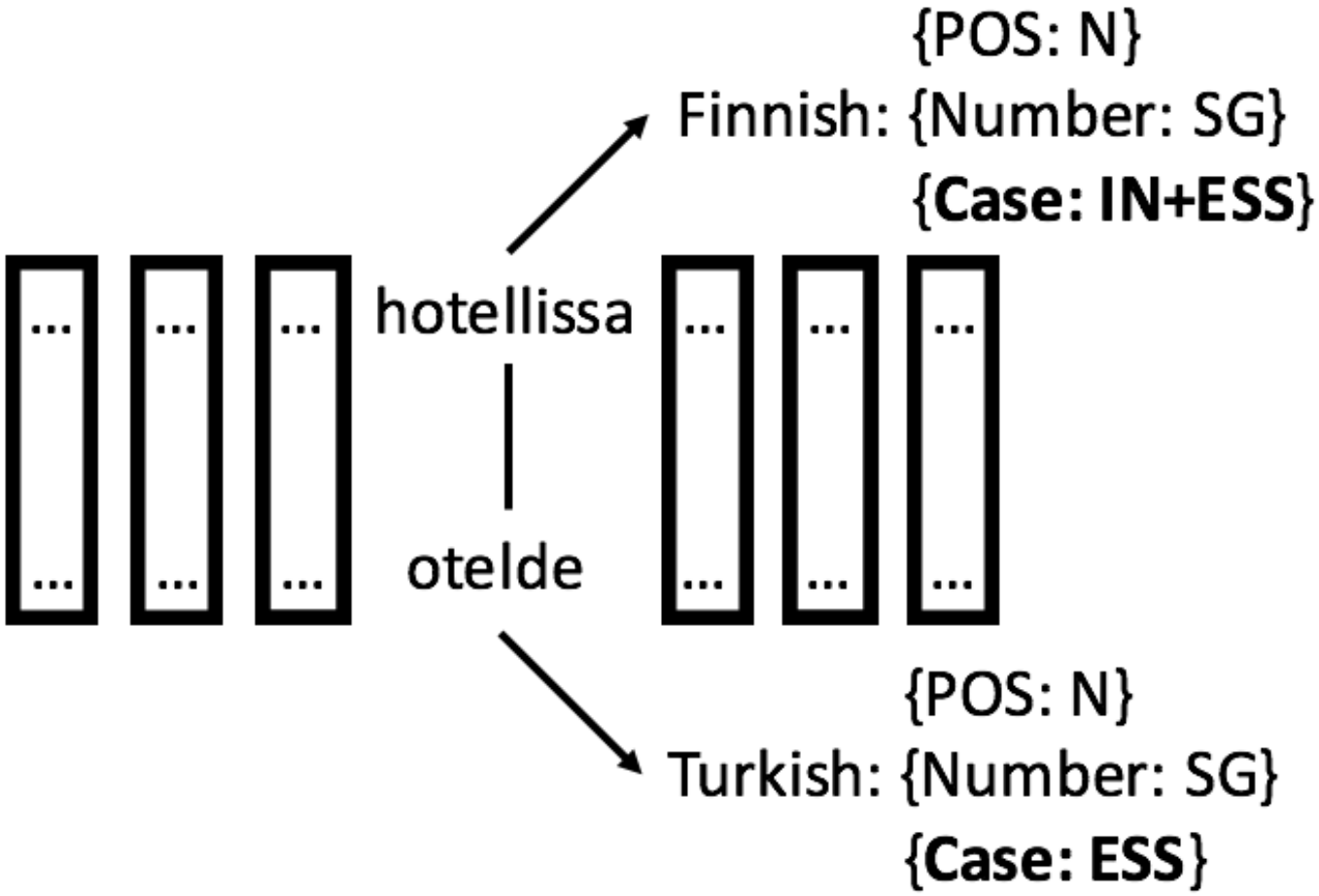
[1] No processes were reported to be G→S

Cross-Lingual Morphological Contrast Mismatches: Methods

- *Overall*: Compared morphological feature specifications in a single category between two aligned words from parallel text.
- Obtained aligned word pairs from OPUS[2] (Tiedemann 2012)
- Word pairs obtained from symmetrized word alignments produced by GIZA++ (Och and Ney 2003) over parallel text from various sources on OPUS, including OpenSubtitles[3] 2012 & 2013, European Medicines Agency (Tiedemann 2009), and EUbookshop (Skadiņš et al. 2014).
- Looked up morphological features of each word in each aligned pair using the UniMorph Wiktionary Corpus.

[2] `http://opus.lingfil.uu.se/`
[3] `http://www.opensubtitles.org/`

Cross-Lingual Morphological Contrast Mismatches: Methods



Cross-Lingual Morphological Contrast Mismatches: Data

- Examined cross-lingual feature contrast mismatches within case and number.
  - Lack of bitext and morphological feature specifications precluded broader examination across other categories which likely have logical hierarchies, including aspect, evidentiality, possession, voice.
- Presenting selected language pairs within case and number:

Case	*Number*
Estonian - Turkish	Italian - Maltese
Finnish - Turkish	Italian - Slovene
Hungarian - Turkish	Russian - Slovene

- Main Results
  - Case comparisons show that specific local cases with 'place' designations (IN, ON, AT) map to more general 'motion' cases (ESS, ALL, ABL; terminology following Radkevich 2010).
  - Number comparisons show that dual maps to plural.

Cross-Lingual Morphological Contrast Mismatches: Case

Finnish - Turkish

↓Fi/Tu→	NOM	GEN	DAT/ALL	ACC	ESS	ABL	X
NOM	1998	36	41	675	16	13	644
ACC	3219	261	84	1425	26	30	1249
GEN	344	238	29	450	6	20	13
PRT	862	31	100	309	4	32	295
IN+ESS	26	8	11	7	303	6	0
ON+ESS	31	21	10	10	93	0	5
IN+ALL	31	16	287	22	14	28	9
ON+ALL	17	4	175	9	0	8	1
IN+ABL	119	28	23	94	24	235	16
ON+ABL	9	4	6	3	5	55	0
X	0	0	0	0	0	0	1483

Cross-Lingual Morphological Contrast Mismatches: Case

Hungarian - Turkish

↓Hu/Tu→	NOM	GEN	DAT/ALL	ACC	ESS	ABL	X
NOM	1444	79	42	173	11	13	585
ACC	515	5	38	466	2	19	69
DAT	15	26	22	10	6	4	4
IN+ESS	10	6	27	0	147	3	5
ON+ESS	9	4	10	10	64	15	1
AT+ESS	2	1	2	2	21	8	0
IN+ALL	3	1	79	0	5	8	1
ON+ALL	22	6	157	20	7	1	1
ALL	6	0	21	4	1	1	1
IN+ABL	1	0	3	11	1	59	1
ON+ABL	14	3	3	10	1	43	1
ABL	5	6	5	7	0	42	0
X	1433	250	43	367	19	25	1470

Cross-Lingual Morphological Contrast Mismatches: Case

Estonian - Turkish

↓Es/Tu→	NOM	GEN	DAT/ALL	ACC	ESS	ABL	X
NOM	401	8	7	106	1	1	133
ACC	222	32	6	66	0	1	103
GEN	358	120	36	195	5	3	8
PRT	362	24	66	268	0	13	23
IN+ESS	16	4	5	4	46	5	1
ON+ESS	11	2	2	5	17	0	0
ESS	1	0	0	0	0	0	0
IN+ALL	0	0	13	1	0	0	0
ALL	5	1	27	5	1	0	1
IN+ABL	5	4	2	1	2	46	0
ABL	2	0	0	0	1	11	0
X	6	0	8	2	0	0	377

Cross-Lingual Morphological Contrast Mismatches: Number

Slovene - Italian

↓Sl/It→	SG	PL	X
SG	3010	777	517
DU	655	1081	417
PL	415	2140	376
X	682	853	334

Maltese - Italian

↓Ma/It→	SG	PL	X
SG	1183	187	462
DU	1	6	0
PL	171	757	81
{SG/PL}	1	5	37

- Dual most frequently maps to plural.

Cross-Lingual Morphological Contrast Mismatches: Number

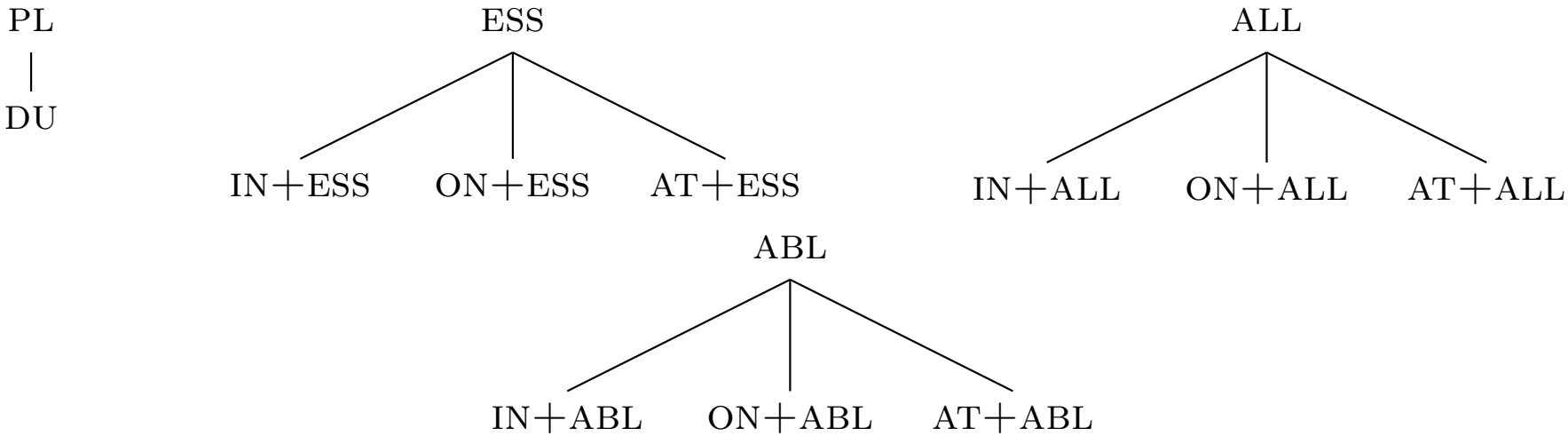
Slovene - Russian

↓Sl/Ru→	SG	PL	X
SG	10550	2767	739
DU	3263 (!)	3027	440
PL	3128	5960	317
X	2805	2712	261

- Slovene-Russian pairing highlights effects of language-specific constructions.
- Russian *dva/dve*, 'two,' calls for (genitive) singular.

Cross-Lingual Morphological Contrast Mismatches: Discussion

- *Expectation*: Exact mapping from specific feature to general feature (e.g. all DU words $\rightarrow$ PL).
- Inexact mappings arise from two main sources of error.
  1. Automatic alignment error
  2. Non-parallel linguistic constructions: Russian-Slovene
- Evidence for following pieces of logical hierarchies:



Cross-Lingual Morphological Contrast Mismatches: Discussion

- In absence of data, theoretical assumptions about relationship of features based on their meaning determines the shape of the hierarchy.
- With data, these assumptions guide analysis.
- *Major issue*: Very little data is available for certain morphological categories which are likely to have complex logical hierarchies, for example: Mood, evidentiality, and switch-reference.

Dependency vs. Logical Hierarchies

- Dependency and logical hierarchies arrive at similar generalizations:
  1. Number relationships discoverable through cross-lingual morphological mismatches.
  2. Example of superclassing in Jingulu (Corbett 2012:22-24) reveals same organization of Harley & Ritter's class features.
- Both offer predictions for which new contrasts may emerge, and what results of loss of contrast may be.
- Dependency hierarchies typically concentrate on morphosyntactic features, not morphosemantic (following distinction by Kibort 2010).
  - Evidence from cross-lingual morphological mismatches alleviates need for observable language-internal effects, allowing logical hierarchies for *morphosemantic* features to be motivated.
- Logical hierarchies offer path for translating features in cases of morphological mismatch when one language is low-resource and it is impossible to discover feature correspondences using automatic methods.

Thank you!

Acknowledgements: Thank you to the CLS 52 anonymous reviewers for very helpful discussions, and thank you to the organizers!

Data and Code: `https://bitbucket.org/rcotterell/cls-code`

Email: jcsg@jhu.edu (John), ryan.cotterell@jhu.edu (Ryan)

References

Clements, George Nick. 1985. The geometry of phonological features. *Phonology Yearbook* 2:225–252.

Corbett, Greville G. 2000. *Number*. Cambridge, UK: Cambridge University Press.

Corbett, Greville G. 2012. *Features*. Cambridge: Cambridge University Press.

Harley, Heidi and Elizabeth Ritter. 2002. Person and number in pronouns: A feature-geometric analysis. *Language* 78(3):482–526.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3):663–687.

Heine, Bernd and Tania Kuteva. 2002. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Jakobson, Roman; Gunnar Fant; and Morris Halle. 1952. *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press.

Kibort, Anna. 2010. Towards a typology of grammatical features. *Features: Perspectives on a Key Notion in Linguistics*, edited by Anna Kibort and Greville G. Corbett, Oxford: Oxford University Press, 64–106.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Radkevich, Nina V. 2010. *On Location: The Structure of Case and Adpositions*. Ph.D. thesis, University of Connecticut, Storrs, CT.

Sagey, Elizabeth. 1990. *The Representation of Features in Non-Linear Phonology: The Articulator Node Hierarchy*. Outstanding Dissertations in Linguistics, New York: Garland.

Skadiņš, Raivis; Jörg Tiedemann; Roberts Rozis; and Daiga Deksne. 2014. Billions of parallel words for free. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland: European Language Resources Association (ELRA).

Sylak-Glassman, John; Christo Kirov; Matt Post; Roger Que; and David Yarowsky. 2015a. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. *Proceedings of the 4th Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, edited by Cerstin Mahlow and Michael Piotrowski, Berlin: Springer, Communications in Computer and Information Science, 72–93.

Sylak-Glassman, John; Christo Kirov; David Yarowsky; and Roger Que. 2015b. A language-independent feature schema for inflectional morphology. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Beijing: Association for Computational Linguistics, 674–680.

Tiedemann, Jörg. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing 5*, edited by N. Nicolov; K. Bontcheva; G. Angelova; and R. Mitkov, Amsterdam: John Benjamins, 237–248.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul: European Language Resources Association (ELRA).