A New Universal Morphological Feature Schema for Rich Morphological Annotation and Cross-Lingual Projection

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, David Yarowsky (PI)

Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD

SFCM
September 17, 2015

Introduction

- Current focus: **Inflectional morphology**
  - High token frequency, all languages use grammatical information it conveys, and it encodes information that is useful to NLP tasks, for example:

Nominal Case	Often correlates with semantic roles
Switch-Reference	Overtly marks cross-clausal NP co-reference
Evidentiality	Encodes speaker's source of information

- Developed a universal morphological feature schema to capture the most basic, fine-grained distinctions made by inflectional morphology across (a large sample of) the world's languages.
- Cross-linguistic validity of features allows schema to function as an 'interlingua' for inflectional morphology, facilitating direct meaning-to-meaning translation.

Universal Morphological Feature Schema: Overview

- Contains 23 *dimensions of meaning*: Morphological categories (e.g. tense, number, case) which contain features that mark distinctions within a common semantic space.
- Over 212 *features*: Represent the most fine-grained distinctions in meaning within each dimension that are conveyed by inflectional morphology in any language.
- Schema allows detailed specification of meaning of inflected words, e.g. Spanish *hablarás* 'you will speak' as:

  speak;V;FIN;IND;POS;DECL;ACT;FUT;2;SG;INFM

  (= speak; VERB; FINITE; INDICATIVE; POSITIVE; DECLARATIVE; ACTIVE; FUTURE; 2ND PERSON; SINGULAR; INFORMAL)

Universal Schema: Construction Methodology

- Surveyed linguistic typology literature to ensure very broad coverage of cross-linguistic diversity, especially low-resource languages.
- *Dimensions of meaning*
  - Identified types of cross-part-of-speech agreement, then searched for dimensions typically expressed on only a single part-of-speech.
- *Features*
  - *Guiding principle*: Features should represent irreducible, "atomic" units of meaning.
  - Allows complex features to be constructed additively, reducing total number of features.
  - For each dimension, found most basic distinctions made by a language.
    - Divisions of scalar property: Number (Sg, Du, Tri, Pauc, Gr. Pauc, Pl)
    - Irreducible orthogonal features: Inverse number (Corbett 2000:161)

Universal Schema: Language-Independent Basis of Features

- Features are defined language-independently.
- *Example*: Aspect defined using Klein's (1994) system, relating time of situation (TSit = { }) to topic time (TT = [ ]). Time of Utterance, TU = |
  - Imperfective —{—[—+++]+++}+++|++ IPFV
  - Perfective —[—{—]—+++}+++|++ PFV
  - Perfect —{—+++}+++[++]+|++ PRF
  - Progressive —{—[—]+++}+++|++ PROG
  - Prospective —[—]—{—+++}+++|++ PROSP
  - Iterative ...[...{—+++}$_{x_1}$...{—+++}$_{x_n}$...]...|... ITER
  - Habitual ...[...{—+++}$_{x_n}$...|...{—+++}$_{x_{n\infty}}$...]... HAB
  - Tense defined similarly, relating TU to TT.
- Language-independent, typologically-informed definitions of features ensure validity of cross-linguistic comparison.
- Universal Morphological Feature Schema does for morphology what Universal Dependencies (Choi et al. 2015) do for syntax, but with finer-grained features specifically for morphology.

Universal Schema: Unique Dimensions

- Schema contains dimensions that are not marked by most other general annotation frameworks.
- Evidentiality: Marks speaker's source of information (direct, hearsay, etc.).
- Switch-Reference: Marks whether an NP in one clause is coreferential with an NP in another clause.
- Information Structure: Marks information as presupposed (topic) or non-presupposed (focus).
- Deixis: Marks distinctions in distance, speaker/addressee reference, visibility, etc. in pronouns.
- Politeness: Typical informal/formal systems (Fr. *tu/vous*), addressee honorifics (e.g. Japanese *teineigo*), bystander honorifics such as Pohnpeian's five levels of honorific speech, and register (e.g. French literary tenses).

Universal Schema: Unique Features

- *Number*: Not only singular, dual, plural, but trial, paucal, greater paucal, as well as greater plural and inverse.
- *Person*: 1st, 2nd, 3rd, as well as 0th (unspecified generic, 'one').
- *Possession*: Type of possession (alienable/inalienable) and detailed characteristics of possessor (person, number, gender, inclusive/exclusive, formal/informal).
- *Case*: Systematic local case features (as in Uralic and Northeast Caucasian languages) informed by global typological survey by Radkevich (2010).

Universal Schema: Full Contents

Dimension	*Features*
Aktionsart	ACCMP, ACH, ACTY, ATEL, DUR, DYN, PCT, SEMEL, STAT, TEL
Animacy	ANIM, HUM, INAN, NHUM
Aspect	HAB, IPFV, ITER, PFV, PRF, PROG, PROSP
Case	ABL, ABS, ACC, ALL, ANTE, APPRX, APUD, AT, AVR, BEN, CIRC, COM, COMPV, DAT, EQU, ERG, ESS, FRML, GEN, INS, IN, INTER, NOM, NOMS, ON, ONHR, ONVR, POST, PRIV, PROL, PROPR, PROX, PRP, PRT, REM, SUB, TERM, VERS, VOC
Comparison	AB, CMPR, EQT, RL, SPRL
Definiteness	DEF, INDEF, NSPEC, SPEC
Deixis	ABV, BEL, DIST, EVEN, MED, NVIS, PROX, REF1, REF2, REM, VIS
Evidentiality	ASSUM, AUD, DRCT, FH, HRSY, INFER, NFH , NVSEN, QUOT, RPRT, SEN
Finiteness	FIN, NFIN
Gender+	BANTU1-23, FEM, MASC, NAKH1-8, NEUT
Info. Structure	FOC, TOP
Interrogativity	DECL, INT
Mood	ADM, AUNPRP, AUPRP, COND, DEB, IMP, IND, INTEN, IRR, LKLY, OBLIG, OPT, PERM, POT, PURP, REAL, SBJV, SIM
Number	DU, GPAUC, GRPL, INVN, PAUC, PL, SG, TRI
Parts of Speech	ADJ, ADP, ADV, ART, AUX, CLF, COMP, CONJ, DET, INTJ, N, NUM, PART, PRO, V, V.CVB, V.MSDR, V.PTCP
Person	0, 1, 2, 3, 4, EXCL, INCL, OBV, PRX
Polarity	NEG, POS
Politeness	AVOID, COL, FOREG, FORM, FORM.ELEV, FORM.HUMB, HIGH, HIGH.ELEV, HIGH.SUPR, INFM, LIT, LOW, POL
Possession	ALN, NALN, PSSD, PSSPNO+
Switch-Reference	CN-R-MN+, DS, DSADV, LOG, OR, SEQMA, SIMMA, SS, SSADV
Tense	1DAY, FUT, HOD, IMMED, PRS, PST, RCT, RMT
Valency	DITR, IMPRS, INTR, TR
Voice	ACFOC, ACT, AGFOC, ANTIP, APPL, BFOC, CAUS, CFOC, DIR, IFOC, INV, LFOC, MID, PASS, PFOC, RECP, REFL

Example 1: Partial Turkish Noun Paradigm

Case	*Definiteness*	*Number*	*Possession*	*Word*	*Gloss*
NOM/ACC	INDEF	SG		ev	'(a) house'
ACC	DEF	SG		evi	'the house'
DAT	*	SG		eve	'to a house'
ESS	*	SG		evde	'in a house'
ABL	*	SG		evden	'from a house'
GEN	*	SG		evin	'of a house'
NOM/ACC	INDEF	SG	PSS1S	evim	'my house' ⟵
NOM/ACC	INDEF	SG	PSS2S	evin	'your house'
NOM/ACC	INDEF	SG	PSS3S	evi	'his/her/its house'
NOM/ACC	INDEF	SG	PSS1P	evimiz	'our house'
NOM/ACC	INDEF	SG	PSS2P	eviniz	'your (pl.) house'
NOM/ACC	INDEF	SG	PSS3P	evleri	'their house'

*Not all dimensions shown

- Can represent as triplets of lemma, inflected word, feature vector: `ev`, `evim`, NOM/ACC;INDEF;SG;PSS1S

Example 2: Hausa 'Completive' Verb Paradigm

Aspect	*Tense*	*Polarity*	*Gender*	*Person*	*Number*	*Word*	*Gloss*
PRF	*	POS	*	1	SG	na tafi	'I went, I {have, had, will have} gone'
PRF	*	POS	MASC	2	SG	ka tafi	'you (m.) went' (etc.)
PRF	*	POS	FEM	2	SG	kin tafi	'you (f.) went'
PRF	*	POS	MASC	3	SG	ya tafi	'he went'
PRF	*	POS	FEM	3	SG	ta tafi	'she went'
PRF	*	POS	*	1	PL	mun tafi	'we went'
PRF	*	POS	*	2	PL	kun tafi	'you all went'
PRF	*	POS	*	3	PL	sun tafi	'they went'
PRF	*	POS	*	0	PL	an tafi	'one went'

*Not all dimensions shown

- Distinguishes the 'zero person': An unspecified, generic participant ('one').

Cross-Lingual Projection of Morphology

- Few-to-none tagged resources for many languages.
- Semantic information relevant to NLP tasks (switch-reference. evidentiality, formality) not overtly marked in languages of interest - e.g., English.
- *Project* tags from high-resource or highly-specified languages to low-resource or underspecified languages.

Cross-Lingual Projection of Morphology

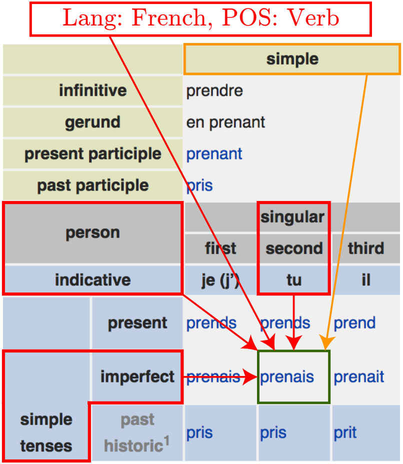
How much noise should we expect from raw, direct cross-lingual projection of morphological features?

- How often will languages that specify the same feature dimension agree?
- Can a consensus of cross-lingual projections provide accurate morphological labels?

Procedure - Wiktionary Extraction and Mapping

- From Wiktionary, extract a database of inflected forms and assign them feature vectors in our schema.
- Wiktionary is a broad-coverage cross-linguistic resource for morphological paradigm data. It is intended to be human-readable, rather than machine-readable, and lacks standardized layouts.

Procedure - Wiktionary Extraction and Mapping

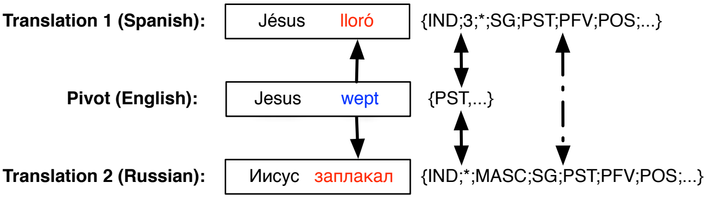


Extracted feature vectors for inflected forms of 883,965 lemmas across 352+ languages in the English edition of Wiktionary. More details in Sylak-Glassman et al. (2015 ACL).

Procedure - Alignment-based Projection

- Use all N and V words in the NT of the NIV English bible as **pivots**.
- Using standard MT tools (Berkeley Aligner), align the English NT to over 800 bibles.
- In Wiktionary, find a feature vector for each foreign word aligned to a pivot. This left 1,683,086 translations covering 47 unique languages across 18 language families.

Example


Translation 1 (Spanish):
Jésus lloró
{IND;3;*;SG;PST;PFV;POS;...}
Pivot (English):
Jesus wept
{PST,...}
Translation 2 (Russian):
Иисус заплакал
{IND;*;MASC;SG;PST;PFV;POS;...}

Agreement Results

- Average pairwise agreement under different genealogical language similarity conditions.

Dimension	Overall	Different Family	Same Family	Same Language
Mood	0.89	0.82	0.95	0.99
Case	0.45	0.23	0.77	0.91
Gender	*0.75*	*0.39*	*0.87*	*0.96*
Number	0.79	0.74	0.88	0.96
Part of Speech	0.74	0.73	0.85	0.94
Person	0.87	0.82	0.93	0.97
Politeness	0.98	0.84	0.99	1.00
Tense	0.73	0.66	0.82	0.95
Voice	0.95	0.83	0.99	0.99
AVERAGE	**0.79**	**0.67**	**0.89**	**0.96**

Evaluating Label Accuracy of Direct Projection

- Evaluate on Wiktionary data in Albanian and Latin.
- Also hold out one aligned language and compare to consensus feature on rest.

Dimension	**Held-Out**	**Albanian**	**Latin**
Case	0.50	0.57	0.81
Gender	0.76	0.74	0.44
Mood	0.91	N/A	0.96
Number	0.83	0.83	0.85
Part of Speech	0.83	0.86	0.59
Tense	0.79	0.84	0.65
Voice	0.95	N/A	0.84
AVERAGE	**0.80**	**0.77**	**0.73**

- The above is a measure of the noise associated with raw direct projection.
- It serves as a baseline for feature accuracy before string and context models.

Conclusion

- Developed typologically-informed, language-independent, very fine-grained morphological feature schema for inflectional morphology.
- Results of projection experiments and systematization of Wiktionary data show that the morphological feature schema already achieves good cross-linguistic coverage and functions well as an interlingua for inflectional morphology.

Thank You!

John Sylak-Glassman	jcsg@jhu.edu
Christo Kirov	ckirov@gmail.com
Matt Post	post@cs.jhu.edu
Roger Que	rque1@jhu.edu
David Yarowsky	yarowsky@jhu.edu

Choi, Jinho; Marie-Catherine de Marneffe; Tim Dozat; Filip Ginter; Yoav Goldberg; Jan Hajič; Christopher Manning; Ryan McDonald; Joakim Nivre; Slav Petrov; Sampo Pyysalo; Natalia Silveira; Reut Tsarfaty; and Dan Zeman. 2015. Universal Dependencies. Accessible at: http://universaldependencies.github.io/docs/.

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

Klein, Wolfgang. 1994. *Time in Language*. New York: Routledge.

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

Sylak-Glassman, John; Christo Kirov; David Yarowsky; and Roger Que. 2015. A language-independent feature schema for inflectional morphology. *Proceedings of the ACL-IJCNLP*, Beijing: Association for Computational Linguistics.

Sylak-Glassman, John; Christo Kirov; Matt Post; Roger Que; and David Yarowsky. To appear. 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*, edited by Michael Piotrowski and Cerstin Mahlow, Berlin: Springer.