The Construction and Evaluation of the LEAFTOP Dataset of Automatically Extracted Nouns in 1480 Languages

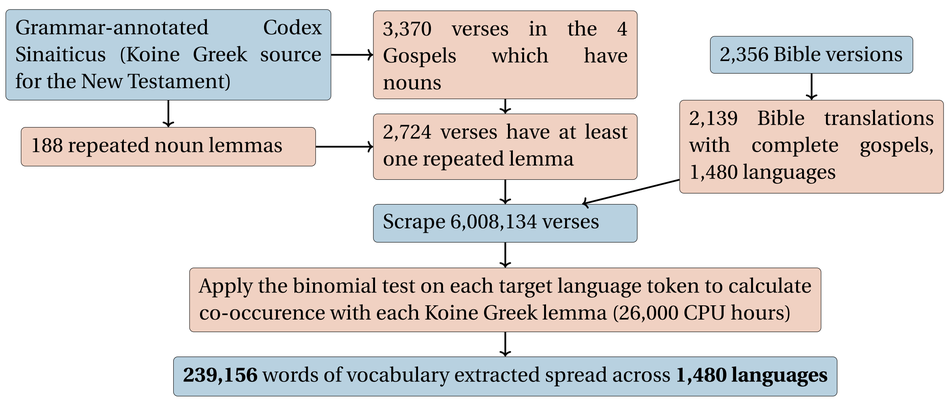
Gregory Baker and Diego Molla

Macquarie University

Overview

This paper discusses a large new dataset with hundreds of high and low-resource languages including both character-based scripts (such as Chinese) and also alphabetic languages. It was created using naive-but-sensible vocabulary extraction techniques (probabilistic word alignment) using a fully-annotated Koine Greek text as a source, with many targets – every language which has all four Gospels translated into it.

Sources and Processing



Example: how does the World English Bible translates the nominative singular of μνημεῖον?

- μνημεῖον appears in the nominative singular only in John 19:41 and John 19:42
- The language (English) is identified as having word break markers, and is alphabetic. Therefore every *unigram* from those verses is analysed.
- There are 2831 verses in the Gospels which either have μνημεῖον in the nominative singular, or don't have μνημεῖον at all.

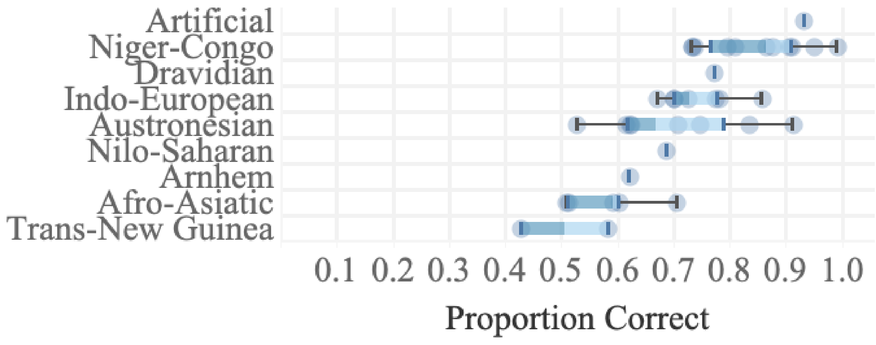
Unigram	Baseline probability of appearance	Appearance count in verses with nom. sing. μνημεῖον	Binomial Test Result
tomb	$\frac{6}{2831} = 2.2 * 10^{-3}$	2	$p = 4.5 * 10^{-6}$
laid	$\frac{29}{2831} = 1.0 * 10^{-2}$	2	$p = 1.04 * 10^{-4}$
garden	$\frac{4}{2831} = 1.4 * 10^{-3}$	1	$p = 2.8 * 10^{-3}$
.	.	.	.
.	.	.	.
the	$\frac{1916}{2831} = 0.68$	2	$p = 0.46$

Best translation is **tomb** with confidence = $\frac{\log(4.5*10^{-6})}{\log(1.04*10^{-4})} = 1.3$

(Full algorithm in the paper)

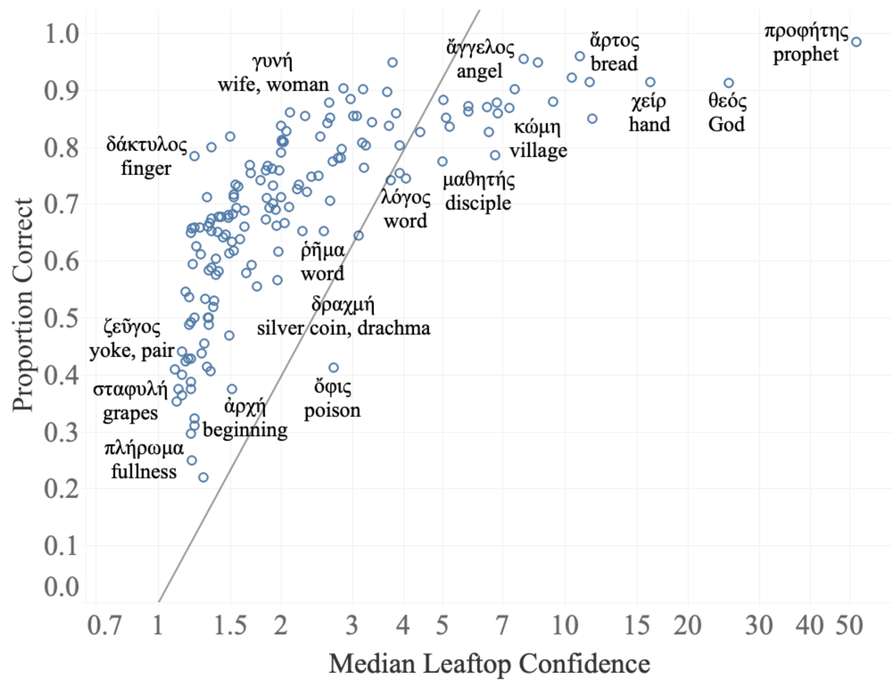
Evaluation

Language Family	Languages Evaluated
Niger-Congo	Fon; Guinea Kpelle; Igbo; Samia; Luganda; Mano; Runyankole; Swahili; Twi; Soga; Yoruba
Afro-Asiatic	Tunisian Arabic; Modern Standard Arabic; Moroccan Arabic; Chadian Arabic
Dravidian	Telugu
Artificial	Esperanto
Arnhem	Gunwinggu
Austronesian	Cebuano; Dobu; Hiligaynon; Hiri Motu; Kilivila; Nyindrou; Takia; Tagalog
Trans-New Guinea	Korafe; Melpa
Indo-European	Bengali; German; French; Hindi; Marathi; Sinhala; Urdu
Nilo-Saharan	Teso



If you know a translator for a language family that's not here, an introduction would be greatly appreciated.

Extraction confidence is correlated with extraction correctness



Proportion Correct = $0.572 * \ln(\text{Median Leaftop Confidence})$

P-value < 0.0001, $R^2 = 0.732$

Other results

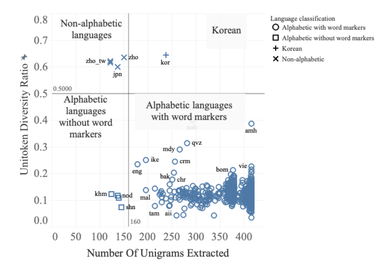


Fig. 1: Orthographic Structure Identification

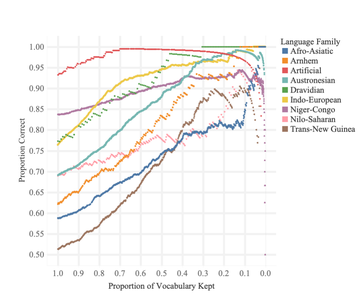
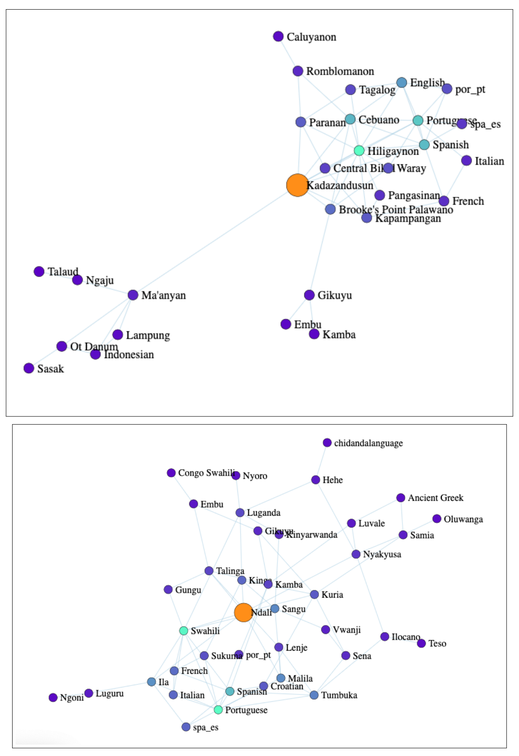


Fig. 2: Trade-offs between vocabulary size and correctness using a threshold for confidence

Data set, including explorer: `https://github.com/solresol/leaftop`

Source code: `https://github.com/solresol/thousand-language-morphology`

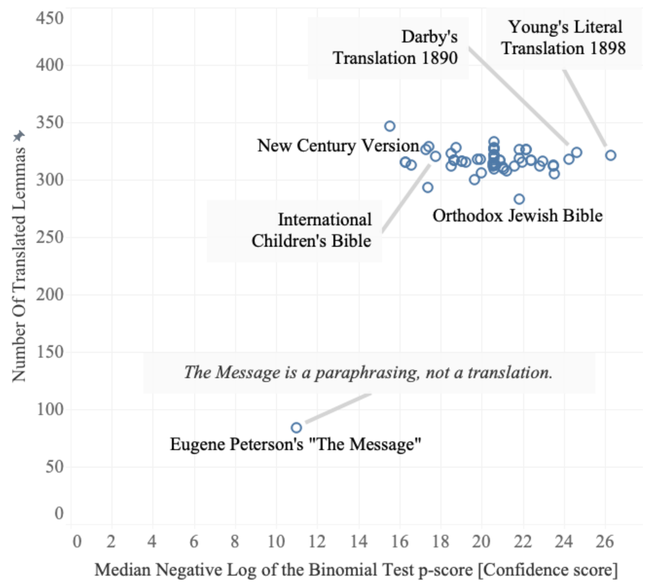
Use case 1: Exploring language similarity



Use case 2: Large-scale grammar morphology in low-resource languages

Lemma	English gloss	Singular in Dobu	Plural in Dobu	Confidence score for singular	Confidence score for plural
ἀδελφός	Brother	tasina	tasinao	1.72	1.3
ὄνομα	Name	esana	esanao	5.9	1.8
προφήτης	Prophet	palopita	palopitao	3.7	4.6
γραμματεύς	Teacher	toe'ita	toe'itao	1.03	1.6
ἄνεμος	Wind	yagila	yagila	2.1	1.6
ὕδωρ	Water	bwasi	ola	2.1	1.4
…					
Lemma	**English gloss**	**Singular in Luganda**	**Plural in Luganda**	**Confidence score for singular**	**Confidence score for plural**
ἀδελφός	Brother	muganda	baganda	2427	1684
ὄνομα	Name	erinnya	amannya	35.6	4.0
προφήτης	Prophet	mulanzi	-	3.6	-
γραμματεύς	Teacher	-	-	-	-
ἄνεμος	Wind	omuyaga	omuyaga	127.1	1.5
ὕδωρ	Water	amazzi	-	14.6	-

Use case 3: Identifying paraphrased translations



The authors would like to thank Daniel Everett and Matías Guzmán Naranjo for their support, and acknowledge the contributions of the freelance translators (some of whom have requested anonymity) who have checked the translations: Benazir Bhagad, Eleanor M. Mendoza, Maureen Y. Ong, Wewalage Roshan Chanaka Perera, Rim Sayed, Ferdaous J., Eric Ojecty, Farah Taymour, Owembabazi Don, Auma Sharot, Okotoi Ruby, Okullo Joel, Bwambale Hamza, Frencelin Laurice, Sidime Amadou, Chimankpa Stanley, Moussa Keita, Paul Malanou, Uhtman Alake and especially Bradley Gewa for his tireless work finding an Australian Aboriginal language translator and all the translators for languages in New Guinea.