Measuring Prefixation and Suffixation in the Languages of the World

Harald Hammarström
Department of Linguistics and Philology
Uppsala University
Box 635
751 26 Uppsala
Sweden
harald.hammarstrom@lingfil.uu.se

Abstract

It has long been recognized that suffixing is more common than prefixing in the languages of the world. More detailed statistics on this tendency are needed to sharpen proposed explanations for this tendency. The classic approach to gathering data on the prefix/suffix preference is for a human to read grammatical descriptions (948 languages), which is time-consuming and involves discretization judgments. In this paper we explore two machine-driven approaches for prefix and suffix statistics which are crude approximations, but have advantages in terms of time and replicability. The first simply searches a large collection of grammatical descriptions for occurrences of the terms 'prefix' and 'suffix' (4 287 languages). The second counts substrings from raw text data in a way indirectly reflecting prefixation and suffixation (1 030 languages, using New Testament translations). The three approaches largely agree in their measurements but there are important theoretical and practical differences. In all measurements, there is an overall preference for suffixation, albeit only slightly, at ratios ranging between 0.51 and 0.68.

1 Introduction

It has long been recognized that suffixing is more common than prefixing in the languages of the world (see Himmelmann 2014, 927 and references therein). More detailed statistics on this tendency are needed to sharpen and evaluate proposed explanations for this tendency. In particular, dense data is needed to properly account for genealogical and areal effects (cf. Murawaki and Yamauchi 2018). With some 7 000 languages in the world, gathering these data is a gargantuan task. In this paper, we investigate three approaches that span the range from minimal to maximal curation.

Motivated by potential functional explanations (Himmelmann, 2014), the ideal measure for prefixing/suffixing would be to count the proportion of prefixes/suffixes per phonological word in a morphologically segmented corpus (cf. Greenberg 1954, 1957). It is believed that such ratios converge as the corpus grows towards infinite amounts of sampled data produced by the speakers of a language, and as such the ratios constitute properties of the language. The ideal measure would range from 0 to (potentially) infinity but, in practice, ratios beyond 5 are unheard of. An alternative equivalent characterization is to have an affixation score (AS) from 0 to (potentially) infinity comprising both prefixes and suffixes, along with a ratio — called the suffix ratio (SR) — from 0.0 to 1.0 of the division of labour between suffixes and prefixes ($\frac{S}{S+P}$). We use this characterization here, remembering that it is only defined for languages which have at least some affixation.

Since large morphologically segmented corpora are not available for a wide range of languages of the world, the ideal token count measure must be approximated. The classic approach, which we may call **Humans read grammars (HRG)**, is for a human to extract the relevant information from grammatical descriptions of the languages of the world. This approach is ideal in many ways, but requires a large amount of manual labour and requires a certain amount of judiciousness on behalf of the curator. While grammars are systematizations of raw text/spoken data, they rarely contain token counts, so this approach can only reflect any specific ratios indirectly. At the other end of the spectrum, a quick-and-dirty approach where **Machines read grammars (MRG)** is possible now that large collections of digitized grammatical descriptions are available and practical to use. We may obtain a crude approximation of the functional load of prefixes/suffixes by simply counting the occurrences of the terms `prefix` and `suffix` of the same grammatical descriptions that were written for a human audience. While there are obvious drawbacks to such a "naive" measure, it has obvi-

Proceedings of the Third Workshop on Computational Typology and Multilingual NLP, pages 81–89

ous advantages in terms of speed, replicability and transparency. A similar crude measure may also be obtained by **Machines Read Raw Text (MRT)** given that a large collection of (not infinite-size, but comparable) raw text collection of New Testaments are available electronically (McCarthy et al., 2020). Correct automatic morphological segmentation and labeling of such a large array languages is not possible at present. Nevertheless, measures inspired by work in Unsupervised Learning of Morphology (Hammarström and Borin, 2011) may be enough to gauge the amount and ratio of affixation even if the tokens cannot be accurately segmented.

2 Related Work

Currently the largest available humanly curated database on prefixation/suffixation in the languages of the world is the WALS chapter 26A by Dryer (2005) featuring 948 languages. It continues a long tradition of growing databases of similar kinds (see, e.g., Himmelmann 2014, 927). We use the Dryer (2005) database here as it represents the culmination of these efforts and is available and methodologically explicit.

Information Extraction from grammatical descriptions has only recently become possible in practice, with the advent of a large collection of digitized grammars (Virk et al., 2020). Given its novelty, only a few embryonic approaches (Virk et al., 2019; Wichmann and Rama, 2019; Macklin-Cordes et al., 2017; Virk et al., 2017; Hammarström et al., 2021) have addressed the task so far. Arguably, the task in the present study is keyword-associated (of the simplest kind) wherefore we follow the method of Hammarström et al. (2021) which requires no tuning of parameters and estimates a noise-level for each source in addition to the simple counts.

While there are no comparable morphologically segmented corpora for a wide range of languages, it should be noted that there is a growing body of scattered resources in the NLP world (e.g., Mott et al. 2020), morphologically segmented texts in the DOBeS and ELAR archives (e.g., Paschen et al. 2020), and Interlinear Glossed Text extracted from miscellaneous publications (see references cited in Round et al. 2020 and Howell 2020). These resources do not yet have the breadth and comparability required for the present study, but the large raw text parallel Bible corpus of McCarthy et al. (2020) does — the culmination of a several decades long tradition of amassing Bible corpora for NLP. Combined with unsupervised morphological segmentation they could provide an excellent resource for direct measurements of affixation. A very large body of work in Unsupervised Learning of Morphology (see Hammarström and Borin 2011 for an overview up to 2010 and, e.g., Eskander et al. 2020 for an overview of more recent work) seeks to do segmentation of raw text. However, despite some progress to date, no off-the-shelf method exists that will segment a very broad range of languages accurately without a large amount of manual tuning of parameters, if even then. Fortunately, for the present task, we only need a score reflecting affixation, not necessarily an accurate segmentation itself. We have thus chosen one of the simplest counting techniques for overrepresentation of initial/terminal string segments (cf. Hammarström and Borin 2011, 322-326) explained in Section 3.3, thought to reflect actual segmentation proportionately. Many other choices would have been possible, with, we suspect, largely equivalent outcomes.

3 Methods

3.1 Humans Read Grammars

Dryer (2005)'s database, reflected in WALS Feature `26A Prefixing vs. Suffixing in Inflectional Morphology`[1], proceeds by calculating a prefix/suffix index for a given language by considering inflectional endings of ten different types, shown in Table 1 (top) along with four example languages. The relative proportion of suffixes versus prefixes ($\frac{S}{S+P}$), called the affixing index (AI), is discretized into five categories along with one category for languages with little or no affixation, as shown in Table 1 (bottom). We only have access to the languages labeled with the discretized labels, not the underlying counts, which would have been a richer rendering (cf. Gerdes et al. 2021). The scope of Dryer (2005) excludes non-inflectional, i.e., derivational prefixes/affixes, pre-/postclitics, intercalated fixes (also known as templatic morphology), tonal changes, preverbs, etc.

3.2 Machines Read Grammars

The data for the experiments in this paper consists of a collection of over 10 000 raw text grammatical descriptions digitally available for computational processing (Virk et al., 2020). A listing of the

[1] Available at online at `https://wals.info/feature/26A`.

		Swedish [swe]		Swahili [swh]		Nuaulu [nxl]	
		P	S	P	S	P	S
(i)	**case affixes on nouns**	-	-	-	-	-	-
(ii)	**pronominal subject affixes on verbs**	-	-	2	-	2	-
(iii)	**tense-aspect affixes on verbs**	-	2	2	-	-	-
(iv)	plural affixes on nouns	-	1	1	-	-	1
(v)	pronominal possessive affixes on nouns	-	-	-	-	0.5	0.5
(vi)	definite or indefinite affixes on nouns	-	1	-	-	-	-
(vii)	pronominal object affixes on verbs	-	-	1	-	-	-
(viii)	negative affixes on verbs	-	-	1	-	-	-
(ix)	interrogative affixes on verbs	-	-	-	-	-	-
(x)	adverbial subordinator affixes on verbs	-	-	-	-	-	-
		0	3	7	0	2.5	1.5
Affixing index (AI)		$\frac{3}{3+0} = 1.0$		$\frac{0}{0+7} = 0.0$		$\frac{1.5}{1.5+2.5} = 0.375$	

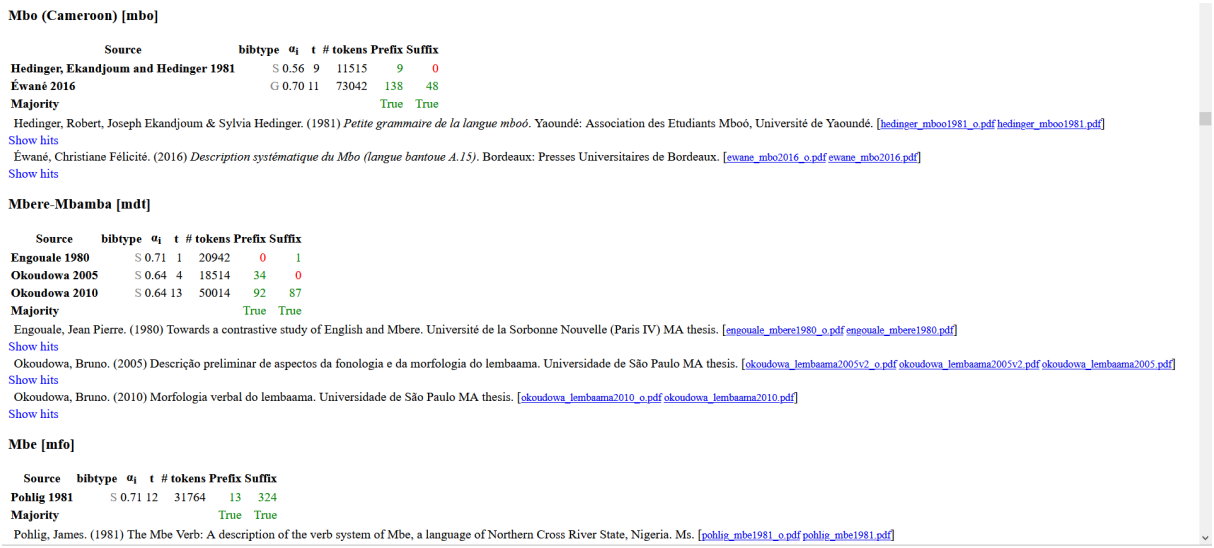
	Label	# lgs	Examples
$P + S \leq 2$	**Little or no inflectional morphology**	141	Thai [tha] (0+0), Vai [vai] (0+2), ...
$0.8 \leq AI$	**Strongly suffixing**	406	Swedish [swe] (3/3), Turkish [tur] (11/11), ...
$0.6 \leq AI < 0.8$	**Weakly suffixing**	123	Beja [bej] (10/13), Mokilese [mkj] (2/3), ...
$0.4 \leq AI < 0.6$	**Equal prefixing and suffixing**	147	Ubykh [uby] (5/10), Kiribati [gil] (2/4), ...
$0.2 \leq AI < 0.4$	**Weakly prefixing**	94	Mohawk [moh] (3/9), Au [avt](1/3), ...
$AI < 0.2$	**Strongly Prefixing**	58	Hunde [hke] (0.5/10), Sango [sag] (0/3), ...
		948	

Table 1: Top: Calculating the affixing index (AI) as per Dryer (2005) given the existence of different types of inflectional prefixes (P) and suffixes (S). The three boldfaced types are considered important enough to count double, hence the 2 points in the respective cells. Bottom: Labels used in Dryer (2005) for different types of prefix/suffix languages given the Affixing index (AI).

collection can be enumerated via the open-access bibliography Glottolog (`glottolog.org`, Hammarström et al. 2020). For each item, we know the (i) language it is written in (the meta-language, usually English, French, German, Spanish, Russian or Mandarin Chinese, see Table 2), (ii) the language(s) described in it (the vernacular, typically one of the thousands of minority languages throughout the world), and (iii) the type of description (comparative study, description of a specific features, phonological description, grammar sketch, full grammar etc). For the experiments in the present study, we used grammars and grammar sketches written in the ten most popular meta-languages. The subset counts 12 032 documents describing 4 287 languages of the world (Table 2). The collection has been OCRed using ABBYY Finereader 14 with using the meta-language as recognition language. The original digital documents are of quality varying from barely legible typescript copies to high-quality scans and even born-digital documents. We have no reason to believe that OCR quality plays any significant role in the experiments to follow. We have however taken care to read latin ligatures accurately as the `fi` ligature (U+FB01) affects the searches for prefix/suffix.

The search over the grammar was done using the Regexps in Table 3 tailored to each language, giving a number of suffix hits S and prefix hits P. In the result output, sources are grouped by language for easy browsing and inspection, as shown in Figure 1. Also included is the total number of tokens[2] of each grammar as well as the "purity level" α_i and associated threshold t automatically calculated using the technique of Hammarström et al. (2021). The suffix ratio for Machines Read Grammars is $SR_{MRG} = \frac{S}{S+P}$ if $S + P > 0$ and conventionally set to 0.5 otherwise.

[2]For Chinese, the Jieba `https://github.com/fxsjy/jieba` tokenizer was employed.



Mbo (Cameroon) [mbo]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Hedinger, Ekandjoum and Hedinger 1981	S	0.56	9	11515	9	0
Éwané 2016	G	0.70	11	73042	138	48
Majority					True	True

Hedinger, Robert, Joseph Ekandjoum & Sylvia Hedinger. (1981) *Petite grammaire de la langue mboó*. Yaoundé: Association des Etudiants Mboó, Université de Yaoundé. [hedinger_mboo1981_o.pdf hedinger_mboo1981.pdf]
Show hits
Éwané, Christiane Félicité. (2016) *Description systématique du Mbo (langue bantoue A.15)*. Bordeaux: Presses Universitaires de Bordeaux. [ewane_mbo2016_o.pdf ewane_mbo2016.pdf]
Show hits

Mbere-Mbamba [mdt]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Engouale 1980	S	0.71	1	20942	0	1
Okoudowa 2005	S	0.64	4	18514	34	0
Okoudowa 2010	S	0.64	13	50014	92	87
Majority					True	True

Engouale, Jean Pierre. (1980) Towards a contrastive study of English and Mbere. Université de la Sorbonne Nouvelle (Paris IV) MA thesis. [engouale_mbere1980_o.pdf engouale_mbere1980.pdf]
Show hits
Okoudowa, Bruno. (2005) Descrição preliminar de aspectos da fonologia e da morfologia do lembaama. Universidade de São Paulo MA thesis. [okoudowa_lembaama2005v2_o.pdf okoudowa_lembaama2005v2.pdf okoudowa_lembaama2005.pdf]
Show hits
Okoudowa, Bruno. (2010) Morfologia verbal do lembaama. Universidade de São Paulo MA thesis. [okoudowa_lembaama2010_o.pdf okoudowa_lembaama2010.pdf]
Show hits

Mbe [mfo]

Source	bibtype	α_i	t	# tokens	Prefix	Suffix
Pohlig 1981	S	0.71	12	31764	13	324
Majority					True	True

Pohlig, James. (1981) The Mbe Verb: A description of the verb system of Mbe, a language of Northern Cross River State, Nigeria. Ms. [pohlig_mbe1981_o.pdf pohlig_mbe1981.pdf]

Figure 1: Example output of the Machines Read Grammars approach.

Meta-language		**# lgs**	**# docs**
English	eng	3 345	7 451
French	fra	792	1 323
German	deu	561	815
Spanish	spa	388	849
Russian	rus	288	537
Mandarin Chinese	cmn	166	249
Portuguese	por	136	285
Indonesian	ind	131	217
Dutch	nld	88	165
Italian	ita	81	139
			12 032

Table 2: Meta-languages of the grammatical descriptions for the present study. The total number of distinct languages covered is 4 287.

3.3 Machines Read Raw Text

New Testament translations for over 1 000 languages are available in the Bible corpus collection of McCarthy et al. (2020). For the purpose of the present study, we assume that whitespace-indicated boundaries correspond to phonological words of the language in question. Languages written in a script that does not indicate word boundaries are excluded from computation. For comparability, we selected only the New Testament and excluded languages which had less than 7 000 verses thereof[3]. The longest text was selected when different versions were available for the same language. A total of 1 030 languages remained.

The type/token ratio is widely taken to be proportionate to the amount of affixation of a language. To measure the division between prefixing and suffixing, we adopt the RA measure of Hammarström (2009, 25-30). As noted above, the technique is one of many variations of the essentially the same theme (Hammarström and Borin, 2011, 322-326). Given any string x and a set W of word types of a corpus, we may calculate the probability of x as final occurrence and the probability of x as a non-final occurrence. $RA(-x)$ is simply the ratio between final and non-final probability, and $RA(x-)$, analogously, the ratio between initial and non-initial probability. For example, $RA(-ing) \approx 35.1$ and $RA(ing-) \approx 0.01$ in the English New Testament. Each segment x may thus be ranked according to prefixhood and suffixhood. From the entire set of attested segments, we keep only the set of suffixes S which are the best suffix-parse (= highest RA) for some word in W and only the set of prefixes P which are the best prefix-parse for some word in W. This makes the very long lists of potential affixes less unwieldy, and the length of the resulting list is believed to be proportionate to the actual number of affixes of each kind. However, it is known that resulting lists of this kind contain segments that are too long compared the actual segmentation, i.e., that contain the true affix plus one or more common characters of the stem or affix of an inner layer. Since we are only interested in the relative amount of prefixation/suffixation here — not the actual segmentation — we may hy-

[3]From inspection of the verse number distribution of the corpus at hand, this number emerges as a cut-off for what may be called a near-complete New Testament.

Heading	**Chinese [cmn]**	**German [deu]**	**English [eng]**	**French [fra]**
Prefix	字首|词头	\W[Pp]r[eä]fix	\W[Pp]refix	\W[Pp]r..?fix
Suffix	后缀|字尾|词尾	\W[Ss]uffix	\W[Ss]uffix	\W[Ss]uffix
Heading	**Italian [ita]**	**Portuguese [por]**	**Russian [rus]**	**Spanish [spa]**
Prefix	\W[Pp]refiss	\W[Pp]refix	\Wпрификс	\W[Pp]refij
Suffix	\W[SS]ufiss	\W[Ss]ufix	\Wсуффикс	\W[Ss]ufij
Heading	**Indonesian [ind]**	**Dutch [nld]**		
Prefix	\W[Pp]refiks|	\W[Pp]refix|		
	\W[Aa]walan	\W[Vv]oorvoegsel		
Suffix	\W[Ss]ufiks|	\W[Ss]uffix|		
	\W[Aa]khiran	\W[Aa]chtervoegsel		

Table 3: Regular expressions for various meta-languages used in the Machines Read Grammar search.

,

	Swedish [swe]		English [eng]		Swahili [swh]	
	x	$RA(x)$	x	$RA(x)$	x	$RA(x)$
1	*-igt*	814.7	*-ned*	556.3	*nili-*	1655.0
2	*-ades*	362.8	*-teth*	475.9	*hawa-*	1365.8
3	*förb-*	343.7	*-ions*	407.9	*wame-*	1341.7
4	*upp-*	316.6	*-nts*	339.9	*-okea*	1261.3
5	*fram-*	248.2	*-ity*	321.4	*walio-*	1140.8
6	*-ligen*	222.7	*-ered*	290.5	*-ieni*	1124.8
7	*förh-*	216.4	*-ied*	284.3	*-zwa*	1108.7
8	*tills-*	203.6	*-neth*	259.6	*nina-*	1100.7
9	*förk-*	203.6	*-tly*	253.4	*wanao-*	1012.3
10	*förm-*	197.3	*-ias*	253.4	*nim-*	988.2
...	...		...		...	

Table 4: Examples of top RA scoring affixes in three languages.

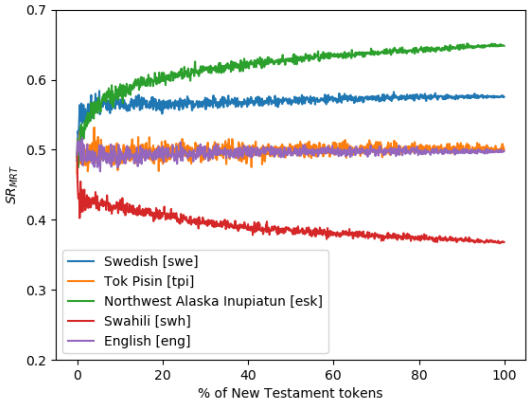


Figure 2: The convergence of SR_{MRT} given increasing percentages of (random) tokens of the New Testament for some example languages including the ones with the lowest (Tok Pisin) and highest (Northwest Alaska Inupiatun) type-token ratio.

pothesize that the erroneous "prolongation" affects prefix and suffix extraction uniformly. Examples of the top RA affixes are shown shown in Table 4 for three languages. The suffix ratio for Machines Read Raw Text is defined to be $SR_{MRT} = \frac{|S|}{|S|+|P|}$. For example $SR_{MRT}(Swedish) = \frac{3629}{3629+2679} \approx 0.58$, $SR_{MRT}(English) = \frac{2930}{2930+2965} \approx 0.5$, $SR_{MRT}(Swahili) = \frac{3109}{3109+5405} \approx 0.37$.

The amount of raw text data needed to reach a stable SR_{MRT} is shown in Figure 2 for some example languages including the most isolating Tok Pisin [tpi] and the record polysynthetic Northwest Alaska Inupiatun [esk]. As expected, all languages show diminishing variation with increased corpus size, but they differ as to how quickly the global value is approximated. Some languages with less morphology appear to reach it with only 10% (or less) of the New Testament, i.e., 700 verses, which corresponds to 15 691 tokens / 2095 types / 63 857 characters in English, 25 239 tokens / 767 types / 95 700 characters in Tok Pisin and 15 792 tokens / 2771 types / 67 745 characters in Swedish. But the more morphologically rich languages appear to require almost the entire text. For the purposes of the present paper, we will assume that the entire New Testament is enough to approximate the true SR_{MRT} of the languages involved.

4 Experiments

4.1 The Individual Measures

For the Humans Read Grammars (HRG) approach, there are no experiments to report, but we may note that the average suffix ratio is $SR_{HRG} = 0.67$ (using the midpoint of the range associated with each label, i.e., 0.1, 0.3, 0.5, 0.7, 0.9) or $SR_{HRG} = 0.65$ if the languages with little affixation are conven-

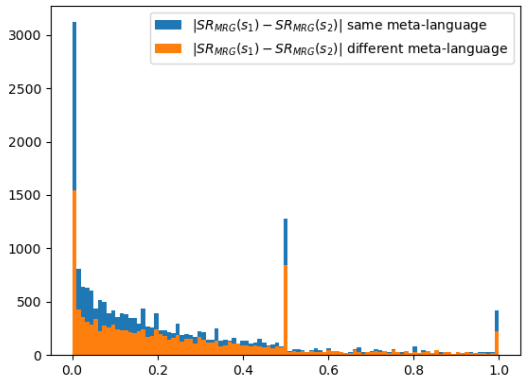


Figure 3: Differences in SR_{MRG} for pairs of different source documents s_1, s_2 describing the *same* language.

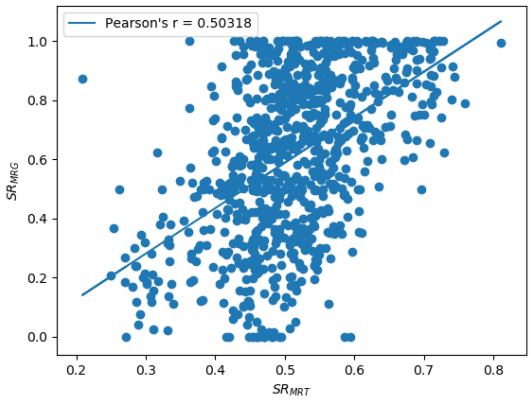


Figure 4: The correlation between MRG and MRT.

tionally said to have a ratio of 0.5.

For the Machines Read Grammars (MRG) approach, there is some latitude in how to treat different sources for the same language. More than half of the languages (2 516 of 4 287) have more than one source and the average number of sources per language is 2.81. Surprisingly, sources for the same language differ quite a lot in their suffix ratio, on average $|SR_{MRG}(s_1) - SR_{MRG}(s_2)| \approx 0.24$ (see Figure 3 for a histogram). This discrepancy is likely not driven by any effects related to different meta-languages as it is ≈ 0.24 when the sources have the same meta-language, only slightly lower than ≈ 0.26 when they do not. Different sources agree on whether $SR_{MRG} > 0.5$ only 68.6% of the time (70.2% if the same meta-language versus 66.3% if different). Manual inspection suggests that the discrepancies are mainly due to differences in scope and attention to functional load across descriptions of the same language, but also relate to differences in author style. For example, Lazard (1981)'s description frequently uses the term 'préfix' along with a hyphenated form *x-*, as expected, but does not use the term suffix when discussing suffixes (that the language does have) which are introduced as *-x* without any explicit accompanying term. The differences notwithstanding, if the suffix ratio of a language is understood as the average suffix ratio of its sources, the average suffix ratio across all 4 287 in MRG is 0.59. It is only a little different, 0.61, if instead we take the source with the most hits (suffix + prefix) per language.

For the Machines Read Raw Text (MRT) approach, the average $SR_{MRT} \approx 0.51$ — only a minimal suffix preference.

4.2 Comparison Between the Three Measures

Table 5 shows a comparison between the three dataset in terms of number of languages in common, average SR for the languages in common, Pearson's r and agreement on whether $SR > 0.5$. HRG and MRG agree on a SR of over 0.6 while MRT exhibits only a small suffix preference. All three measures are correlated with an $r > 0.5$. A scatter plot for $MRG \cap MRT$ — the two continuous measures — is shown in Figure 4. The agreement between all three measures increases to around 0.7 if we only consider the polarity of SR.

We should not expect these measures to fully agree given the significant theoretical differences. HRG has been forcibly discretized, considers only inflectional morphology and has an opaque link to the token ratio. MRG is quite sensitive to the descriptive aims (and whims) of particular authors and is unable to discern the type and context of affixes. Some authors discuss more comparative aspects, some include detailed discussions of morphophonology, some describe subordinate clauses in more detail than others and so on. It is telling that MRG agrees with the other measures roughly as much as MRG for different sources of the same language. It thus seems that this accuracy is a natural limit to what naive keyword counting can achieve on this (and similar) tasks. Similarly, MRT can not differentiate between derivational, inflectional or fossilized/productive affixation and it is not known how close the HRT measure is to the ideal token count and/or if there is a simple improvement.

To exemplify these differences, consider the comparison of SR-measurements for ten randomly

Dataset	# lgs	Avg SR	Pearson's r	SR polarity agreement
MRT ∩ HRG	306	MRT: 0.53, HRG: 0.68	0.54	0.73
MRT ∩ MRG	880	MRT: 0.51, HRG: 0.61	0.50	0.67
HRG ∩ MRG	917	HRG: 0.65, MRG: 0.65	0.67	0.75
∩ All three	301	MRT: 0.53, HRG: 0.66, MRG: 0.64	-	-

Table 5: Overlap and comparison of the three approaches for measuring the suffix ratio.

Language		$\|S\|$	$\|P\|$	Tokens	Types	SR_{MRT}	SR_{HRG}	SR_{MRG}
Adamawa Fulfulde	fub	2639	1773	138713	8394	0.60	0.70	0.91
Alekano	gah	2524	3879	206212	14206	0.39	0.90	0.85
Amharic	amh	7158	6259	99866	24751	0.53	0.70	0.67
Burarra	bvr	811	674	120804	1544	0.55	0.30	0.25
Nogai	nog	5876	2509	127036	18787	0.70	0.90	0.75
Nyankole	nyn	3007	6780	109603	19855	0.31	0.10	0.14
Páez	pbb	2588	1646	97749	8043	0.61	0.90	0.67
Uighur	uig	5908	1869	140666	20655	0.76	0.90	0.79
Woun Meu	noa	3262	1562	217057	10167	0.68	0.90	0.91
Wubuy	nuy	814	775	69363	2172	0.51	0.50	0.38

Table 6: New Testament data size and SR-measurements for ten randomly chosen languages featured for all three methods.

chosen languages in Table 6. We have not been able to investigate in depth the judgment of MRT of Alekano as a prefix-dominant language. A possibility informally observed in some other cases is that frequent stems are judged as prefixes. Indeed the MRT method lacks any information needed to distinguish stems from affixes if not for their frequency distributions. Amharic is written in an abugida script which should theoretically make the MRT estimate more coarse grained, and this is possibly reflected in its comparatively lower SR_{MRT}. Burarra is judged by MRT as a suffixing language, but here the explanation may be related to the orthography. The Burarra words as rendered in the Bible corpus contain a lot of dashes, likely indicating (some? all?) affix boundaries, possibly interfering with the MRT method (but this has not been investigated in depth). The two grammars used in MRG for Wubuy (one of which, Heath 1984, also underlies the HRG value) do discuss the prefixes much more than the suffixes since the prefix system indicating noun classes in this language is quite complicated.

Judging from the three-way comparison, the MRT measure is more often deviant from the other two. A closer look is needed to determine the source(s) of discrepancy more systematically. More research is needed into the robustness of the MRT-measure and related techniques, especially as it concerns the influence of orthography/writing system.

While the above discusion concerns the division of labour between suffixes and prefixes, we should also note how well the amount of affixation can be measured. In HRG, 141 of 948 languages are said to have "Little Affixation". Simple logistic regression gives an accuracy of 86% in predicting this class from the type/token ratio of MRT and 85% in predicting the class from the suffix, prefix, purity level and token count of the grammar with the most hits for each language. But these numbers do not improve on the baseline, and so add no actual information as to this class. Furthermore, there is only a weak correlation ($r \approx 0.15$) between the type-token ratio of MRT and the ratio of affixation hits to tokens times purity level. Clearly, predicting the amount of affixation is not as simple as it appears at first glance (cf. Bentz et al. 2016).

5 Conclusion

We have compared three ways to obtain data on the amount of prefixes/suffixes in the languages of the world. The three measures, correlate to a high degree but none can be said to reflect an ideal measure. At the same time, there are considerable differences in the measurements of individual languages. These differences reflect differences in aim and scope as well as sketchy measurements. The

Humans Read Grammars method only focusses on inflectional morphology with only weak integration of functional load. The Machines Read Grammars approach is vulnerable to differences in scope of description and individual styles, of which there is plenty of variation for the same language. More research is needed in to see to what extent these dimensions of variation can somehow be normalized automatically. The Machines Read Raw Text method reads a very noisy reflection of prefixation/suffixing from the raw data and cannot differentiate between derivational, inflectional or fossilized/productive affixation. The simple measure used here should be abandoned in favour of a more complicated, but less noisy measure. The resulting database, in total spanning the tremendous 4 437 languages, is freely available for future research at Zenodo http://doi.org/10.5281/zenodo.4731249 on a Creative Commons Attribution 4.0 International license.

Acknowledgements

This research was made possible thanks to the financial support of the From Dust to Dawn: Multilingual Grammar Extraction from Grammars project funded by Stiftelsen Marcus och Amalia Wallenbergs Minnesfond 2017.0105 awarded to Harald Hammarström (Uppsala University) and the Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage (DReaM) Project awarded 2018-2020 by the Joint Programming Initiative in Cultural Heritage and Global Change, Digital Heritage and Riksantikvarieämbetet, Sweden.

References

Christian Bentz, D. Alikaniotis, T. Samardžić, and P. Buttery. 2016. Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*, 23(2):1–41.

Matthew S. Dryer. 2005. Prefixing versus suffixing in inflectional morphology. In Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath, editors, *World Atlas of Language Structures*, pages 110–113. Oxford: Oxford University Press.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans, and Smaranda Muresan. 2020. Morphagram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7114–7124, Marseille, France. European Language Resources Association.

Kim Gerdes, Sylvain Kahane, and Xinying Chen. 2021. Typometrics: From implicational to quantitative universals in word order typology. *Glossa: a journal of general linguistics*, 6(1):1–31.

Joseph H. Greenberg. 1954. A quantitative approach to the morphological typology of language. In Robert F. Spencer, editor, *Method and Perspective in Anthropology: Papers in Honor of Wilson D. Wallis*, pages 192–220. Minneapolis: University of Minnesota Press.

Joseph H. Greenberg. 1957. Order of affixing: A study in general linguistics. In Joseph H. Greenberg, editor, *Essays in linguistics*, pages 86–97. Chicago: University of Chicago Press.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. Glottolog 4.3. Jena: Max Planck Institute for the Science of Human History. Available at http://glottolog.org. Accessed on 2020-11-02.

Harald Hammarström. 2009. *Unsupervised Learning of Morphology and the Languages of the World*. Ph.D. thesis, Chalmers University of Technology and University of Gothenburg.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Harald Hammarström, One-Soon Her, and Marc Tang. 2021. Keyword spotting: A quick-and-dirty method for extracting typological features of language from grammatical descriptions. In *Proceedings of SLTC 2020*, page submitted, Gothenburg, Sweden. Nordic Journal of Language Technology.

Jeffrey Heath. 1984. *Functional Grammar of Nunggubuyu*. Canberra: Australian Institute of Aboriginal Studies.

Nikolaus Himmelmann. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language*, 90(4):927–960.

Kristen Howell. 2020. *Inferring Grammars from Interlinear Glossed Text: Extracting Typological and Lexical Properties for the Automatic Generation of HPSG Grammars*. Ph.D. thesis, University of Washington.

Gilbert Lazard. 1981. Le dialecte des juifs de kerman. In *Monumentum Georg Morgenstierne 1*, volume 21 of *Acta Iranica*, pages 333–346. Paris: Brill.

Jayden L. Macklin-Cordes, Nathaniel L. Blackbourne, Thomas J. Bott, Jacqueline Cook, T. Mark Ellison, Jordan Hollis, Edith E. Kirlew, Genevieve C. Richards, Sanle Zhao, and Erich R. Round. 2017.

Robots who read grammars. Poster presented at Co-EDL Fest 2017, Alexandra Park Conference Centre, Alexandra Headlands, QLD.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2877–2885, Marseille, France. European Language Resources Association.

Justin Mott, Ann Bies, Stephanie Strassel, Jordan Kodner, Caitlin Richter, Hongzhi Xu, and Mitchell Marcus. 2020. Morphological segmentation for low resource languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3989–3995, Marseille, France. European Language Resources Association.

Yugo Murawaki and Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution*, 3(1):13–25.

Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (doreco). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2650–2659, Marseille, France. European Language Resources Association.

Erich Round, Mark Ellison, Jayden Macklin-Cordes, and Sacha Beniamine. 2020. Automated parsing of interlinear glossed text from page images of grammatical descriptions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2871–2876, Marseille, France. European Language Resources Association.

Shafqat Mumtaz Virk, Lars Borin, Anju Saxena, and Harald Hammarström. 2017. Automatic extraction of typological linguistic features from descriptive grammars. In Kamil Ekštein and Václav Matoušek, editors, *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings*, volume 10415 of *Lecture Notes in Computer Science*, pages 111–119. Berlin: Springer.

Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, and Søren Wichmann. 2020. The dream corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 871–877. Marseille, France: European Language Resources Association, Marseille, France.

Shafqat Mumtaz Virk, Azam Sheikh Muhammad, Lars Borin, Muhammad Irfan Aslam, Saania Iqbal, and Nazia Khurram. 2019. Exploiting frame-semantics and frame-semantic parsing for automatic extraction of typological information from descriptive grammars of natural languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, page 1247–1256. Varna, Bulagaria: NCOMA Ltd.

Søren Wichmann and Taraka Rama. 2019. Towards unsupervised extraction of linguistic typological features from language descriptions. First Workshop on Typology for Polyglot NLP, Florence, Aug. 1, 2019 (Co-located with ACL, July 28-Aug. 2, 2019).