

# A Survey of Computational Morphological Resources for Low-Density Languages

Harald Hammarström

December 27, 2008

## 1 Introduction

The present paper looks at computational morphological resources for so-called low-density languages. First, we give a definition of “low-density language” in terms of the economic power of its speakers. Next, we survey which low-density languages have (a published description of) a computational morphological analyser of some sort.

A survey of this kind is relevant for understanding how language resources come about [for low-density languages], which in turn is relevant for the broader questions of language survival [of low-density languages]. It is also relevant for the design and application of unsupervised NLP tools, which potentially offer their highest added value for low-density languages. These issues have been discussed in more detail in several works (Yli-Jyrä, 2005; Streiter, Scannell, & Stuflessen, 2006; Berment, 2004; David & Maxwell, 2008; Sornlertlamvanich, 2008; A. K. Singh, 2008) but there is so far no empirically grounded understanding of how language resources for low-density languages emerge. For this reason, in this paper we focus on a comprehensive survey, with aim that of a better understanding of the full picture in the future.

We also aim to gather together work which is otherwise scattered in very disparate fora to facilitate comparison – presumably important lessons from one language can be taught to another language in a similar situation.

We focus on resources for morphological analysis since it is one of the bottom layers<sup>1</sup> of analysis of written language data. The layer below, namely

---

<sup>1</sup>As per the language resources pyramid (usually presented in 2D, like a triangle) given

raw text data, is addressed by Baldwin, Bird, and Hughes (2006). It is clear that the languages for which raw text data appears on the web<sup>2</sup> is much larger than those for which there is a published description of a morphological analyser.

## 2 Low-Density Languages

For this survey to be practical we need a clear definition of what a low-density language is, rather than a sliding scale or a prototypical set of properties.

We suggest that a low-density language may be characterized in terms of the economic power of its speakers. Hence, in analogy with GDP, we define the Gross Language Product (GLP), of a language as the total market value of all final goods and services produced by the speakers of the language within a calendar year. Since there is no detailed data available to compute this statistic accurately for a large number of languages, we use the following formula to estimate GLP:

$$GLP(L) = S(L) \cdot GDP\text{-per-capita}(Country(L))$$

In other words, we employ the convenient fiction that each language  $L$  has a principal country  $Country(L)$  and then pretend that the  $GDP$ -per-capita for that country is indicative also for the  $S(L)$  (first language) speakers of  $L$  in that country. Data for GDP per capita from Central Intelligence Agency of the United States (2007) and speaker numbers from Gordon (2005) are accessible. Table 1-2 shows the top 100 densent languages according to this metric.

Some comments are in order.

- The GDP-figures used are not PPP-adjusted since the prices for NLP related services appear to have little to do with local prices for basic commodities.
- Ideally, one would like to count second language speakers along with first language speakers. This would give a much better estimate, especially of the density of dead languages. Unfortunately, such data are not systematically available for a wide range of languages.

---

in Lars Borin's lectures, cf. Beesley (2004) and Berment (2004, 18-26).

<sup>2</sup>According to a popular article *Weaving a Web of linguistic diversity*, <http://www.guardian.co.uk/GWeekly/Story/0,3939,427939,00.html>, 2001-01-25, retrieved 2006-09-12, this number is about 1 000. Though it is not clear how this figure was computed.

#	Language	iso-639-3	GLP	Pop.	GDP-per-capita
1	English	eng	14112019141500	309297750	45626
2	Spanish	spa	11466115307496	322299171	35576
3	Japanese	jpn	4210405702398	122388399	34402
4	German, Standard	deu	3845767908070	95392978	40315
5	Portuguese	por	3723229093580	177457180	20981
6	French	fra	2606363422200	64834911	40200
7	Italian	ita	2207898410784	60989984	36201
8	Chinese, Mandarin	cmn	2146742158782	873014298	2459
9	Russian	rus	2146466954800	145031551	14800
10	Korean	kor	1328565491640	66977490	19836
11	Dutch	nld	805812974253	17370777	46389
12	Turkish	tur	471145207462	50535794	9323
13	Polish	pol	465485547296	42658133	10912
14	Swedish	swe	443139531525	8789835	50415
15	Greek	ell	360217197900	12258540	29385
16	Bavarian	bar	349629329322	7667478	45599
17	Schwyzerdütsch	gsw	339134884000	6044000	56111
18	Lombard	lmo	330654684855	9133855	36201
19	Danish	dan	302298082240	5299756	57040
20	Napoletano-Calabrese	nap	255122891199	7047399	36201
21	Finnish	fin	244729455832	5232728	46769
22	Catalan-Valencian-Balear	cat	237196860928	6667328	35576
23	Czech	ces	197516975282	11525089	17138
24	Chinese, Wu	wuu	189773325000	77175000	2459
25	Hungarian	hun	189214851600	13611600	13901
26	Hindi	hin	175884141643	180764791	973
27	Sicilian	sen	174942056520	4832520	36201
28	Romanian	ron	173246830884	23248367	7452
29	Javanese	jav	139312813500	75508300	1845
30	Chinese, Yue	yue	134779260482	54810598	2459
31	Arabic, Najdi Spoken	ars	134439777600	9863520	13630
32	Malay	mly	132190335777	17604253	7509
33	Ukrainian	ukr	119705990470	39441842	3035
34	Hebrew	heb	117078855000	5055000	23161
35	Chinese, Min Nan	nan	113674565935	46227965	2459
36	Galician	glg	113430518400	3188400	35576
37	Piemontese	pms	112462750620	3106620	36201
38	Chinese, Jinyu	cjy	110655000000	45000000	2459
39	Azerbaijani, South	azb	109564908000	24364000	4497
40	Farsi, Western	pes	109237171137	24291121	4497
41	Tswana	tsn	91581075720	4407174	20780
42	Chinese, Xiang	hsn	88560885000	36015000	2459
43	Kurdish, Northern	kmr	84191398115	9030505	9323
44	Arabic, Algerian Spoken	arq	83227665000	21097000	3945
45	Bengali	ben	82284767162	171070202	481
46	Arabic, Hijazi Spoken	acw	81780000000	6000000	13630
47	Saxon, Upper	sxu	80630000000	2000000	40315
48	Venetian	vec	78932189787	2180387	36201
49	Thai	tha	76388430912	20229987	3776
50	Arabic, Egyptian Spoken	arz	73727112000	46311000	1592
51	Chinese, Hakka	hak	73617441181	29937959	2459
52	Emiliano-Romagnolo	eml	73130074512	2020112	36201
53	Yiddish, Eastern	ydd	72784832160	3142560	23161
54	Croatian	hrv	71033369490	6214643	11430
55	Limbürgisch	lim	69583500000	1500000	46389
56	Ligurian	lij	69536654649	1920849	36201
57	Slovak	slk	68983077920	5011120	13766
58	Telugu	tel	67806694494	69688278	973
59	Marathi	mar	66212442751	68049787	973
60	Tamil	tam	64237654600	66020200	973
61	Thai, Northeastern	tts	56640000000	15000000	3776
62	Zulu	zul	55879074746	9563422	5843
63	Kazakh	kaz	55542767289	8178879	6791
64	Vietnamese	vie	55318273119	67379139	821
65	Sardinian, Logudorese	src	54301500000	1500000	36201
66	Auvergnat	auv	52863000000	1315000	40200
67	Vlaams	vls	52464896000	1202000	43648
68	Punjabi, Western	pnb	51629466957	60812093	849
69	Urdu	urd	51367538571	60503579	849
70	Chinese, Gan	gan	50606220000	20580000	2459
71	Sunda	sun	49815000000	27000000	1845
72	Walloon	wln	48885760000	1120000	43648
73	Bulgarian	bul	48436572699	8954811	5409
74	French, Cajun	frc	45959000000	1000000	45959
75	Serbian	srp	45739846348	11139758	4106

Table 1: 140 top-density languages, ranked by GLP [Page 1(2)].

#	Language	iso-639-3	GLP	Pop.	GDP-per-capita
76	Slovenian	slv	45518829850	1984775	22934
77	Gujarati	guj	44861270328	46106136	973
78	Indonesian	ind	42699488130	23143354	1845
79	Arabic, Libyan Spoken	ayl	42581260000	4505000	9452
80	Arabic, Moroccan Spoken	ary	42292382600	19480600	2171
81	Xhosa	xho	42152091474	7214118	5843
82	Belarusan	bel	41809393608	9081102	4604
83	Luxembourgish	ltz	40800831336	390618	104452
84	Afrikaans	afr	34858630997	5965879	5843
85	Malayalam	mal	34791658300	35757100	973
86	Kannada	kan	34391658000	35346000	973
87	Guadeloupean Creole French	gcf	34107850800	848454	40200
88	Okinawan, Central	ryu	33861372570	984285	34402
89	Turkmen	tuk	33810654240	6403533	5280
90	Lithuanian	lit	33521764006	3125281	10726
91	Arabic, Mesopotamian Spoken	acm	330086000000	15100000	2186
92	Frisian, Western	fri	32472300000	700000	46389
93	Arabic, Tunisian Spoken	aeb	31488759000	9247800	3405
94	Oriya	ori	30842673582	31698534	973
95	Cebuano	ceb	30666558060	20043502	1530
96	Arabic, Sa'idi Spoken	aec	30088800000	18900000	1592
97	Friulian	fur	28743594000	794000	36201
98	Hausa	hau	28124568000	24162000	1164
99	Arabic, North Levantine Spoken	apc	27975144835	14309537	1955
100	Bashkir	bak	27696468400	1871383	14800
101	Hawai'i Creole English	hwc	27575400000	600000	45959
102	Gronings	gos	27462288000	592000	46389
103	Punjabi, Eastern	pan	27250522992	28006704	973
104	Azerbaijani, North	azj	27228603353	7059529	3857
105	Chuvash	chv	27149031200	1834394	14800
106	Bhojpuri	bho	25874537528	26592536	973
107	Chinese, Min Bei	mnp	25312946000	10294000	2459
108	Madura	mad	25267090500	13694900	1845
109	Zhuang, Northern	ccx	24590000000	10000000	2459
110	Welsh	cym	24467307508	536258	45626
111	Tagalog	tgl	24327149940	15900098	1530
112	Maithili	mai	24128047286	24797582	973
113	Réunion Creole French	rcf	24120000000	600000	40200
114	Tatar	tat	23828473600	1610032	14800
115	Thai, Northern	nod	22691479296	6009396	3776
116	Yoruba	yor	22496628000	19327000	1164
117	Chinese, Min Dong	edo	22384663063	9103157	2459
118	Gaelic, Irish	gle	22341570000	355000	62934
119	Arabic, Sudanese Spoken	apd	22251592000	18986000	1172
120	Sotho, Northern	nso	21675876431	3709717	5843
121	Breton	bre	21415424400	532722	40200
122	Igbo	ibo	20952000000	18000000	1164
123	Basque	eus	20922530208	588108	35576
124	Umbundu	umb	20026408640	4002880	5003
125	Awadhi	awa	20005603912	20560744	973
126	Tsonga	tso	19136438515	3275105	5843
127	Sinhala	sin	18957847104	13220256	1434
128	Thai, Southern	sou	18880000000	5000000	3776
129	Uyghur	uig	18691918829	7601431	2459
130	Latvian	lav	18677424712	1543844	12098
131	Sindhi	snd	18136338000	21362000	849
132	Armenian	hye	18040062720	6723840	2683
133	Plautdietsch	pdt	17465069122	401699	43478
134	Estonian	est	17391861987	1075497	16171
135	Arabic, South Levantine Spoken	ajp	16253525000	6145000	2645
136	Corsican	cos	16160400000	402000	40200
137	Icelandic	isl	15882232320	239768	66240
138	Uzbek, Northern	uzn	15487566984	18795591	824
139	Mbundu	kmb	15009000000	3000000	5003
140	Kabardian	kbd	14977600000	1012000	14800
141	Assamese	asm	14958902000	15374000	973
...					

Table 2: 140 top-density languages, ranked by GLP [Page 2(2)].

- Note that we define a low-density language as one whose speakers have low GLP. It is then an empirical question whether low-density languages actually turn out to have low amounts of NLP infrastructure.<sup>3</sup>

For the purposes of this paper, we will put the threshold of low-density at 100 billion dollars of GLP. This threshold was chosen based on nothing more than its large number of zeroes and the fact that a convenient number of non-low-density languages emerge. With this setting, there are currently 40 non-low-density languages, and all the rest, beginning with Tswana at rank #41, are low-density languages<sup>4</sup>

Also, as for non-low-density languages, it is obvious that all<sup>5</sup> of them have fair amounts of NLP infrastructure except for an important class of languages with the following properties:

- They are not popularly written
- In the country where they are spoken, there is a standardized close relative which is the preferred language for written communication

The methodology used for gathering the items in the survey was general knowledge, browsing of the meta-literature, the corpora-list and googling suitable for each of the 100 densest low-density languages.

For languages which have very little morphology we listed some other NLP work at a comparable stage.

## 3 Survey

### 3.1 Europe

#### 3.1.1 Irish

- Dhonnchadha, Pháidín, and Genabith (2003)
- Sulger (2008)

---

<sup>3</sup>Another possibility would have been to reserve the label low-density for actual low-density of NLP resources and use the label low-power (or the like) for the GLP-measure used here. We do not opt for this choice as it seems the term low-density language is currently used with the meaning closer to low-power than actual low-density of NLP infrastructure. For other choices of terminology see Streiter et al. (2006, 2-3).

<sup>4</sup>Of course, the division of languages is always debatable.

<sup>5</sup>The only one I am not immediately convinced of is South Azerbaijani.

### **3.1.2 Welsh**

- Chrupała (2008)

### **3.1.3 Czech**

- Chrupała (2008)
- Schmid and Laws (2008)

### **3.1.4 Slovene**

- Erjavec and Džeroski (2004)
- Chrupała (2008)
- Hajič (2000)

### **3.1.5 Latvian**

- Paikens (2007)

### **3.1.6 Lithuanian**

- Rimkutė, Daudaravičius, and Utkā (2007)

### **3.1.7 Ancient Greek**

- Lee (2008)

### **3.1.8 Basque**

- Alegria, Artola, Sarasola, and Urkia (1996)

### **3.1.9 Sámi**

- Trosterud (2008d)

### **3.1.10 Estonian**

- Uibo (2002)
- Kaalep and Vaino (2001)
- Müürisep et al. (2003)

### **3.1.11 Faroese**

- Trosterud (2008a)

### **3.1.12 Icelandic**

- Loftsson (2008a)
- Loftsson (2008b)

### **3.1.13 Bulgarian**

- Slavcheva (2003)
- Simov, Osenova, Kolkovska, Balabanova, and Doikoff (2004)

### **3.1.14 Serbian**

- Krstev, Vitas, and Erjavec (2004)

### **3.1.15 Mordvin**

- Prószéky and Novák (2005)

### **3.1.16 Udmurt**

- Prószéky and Novák (2005)

### **3.1.17 Komi**

- Prószéky and Novák (2005)

### **3.1.18 Mansi**

- Prószéky and Novák (2005)

### **3.1.19 Khanty**

- Prószéky and Novák (2005)

### **3.1.20 Tundra Nenets**

- Prószéky and Novák (2005)

### **3.1.21 Nganasan**

- Prószéky and Novák (2005)

### **3.1.22 Latin**

- Forsberg (2007)

## **3.2 Asia**

### **3.2.1 Malayalam**

- Idicula and David (2007)

### **3.2.2 Kannada**

- Vikram and Urs (2007)

### **3.2.3 Sinhala**

- Herath, Ikeda, Yokoyama, Isahara, and Ishizaki (1989)

### **3.2.4 Tamil**

- Anandan, Parthasarathi, and Geetha (2002)
- Viswanathan, Kumar, Shanmugam, and Arulmozi (2003)

### **3.2.5 Bengali**

- Dasgupta and Ng. (2007)
- Sarkar and Bandyopadhyay (2008)
- Dandapat, Sarkar, and Basu (2007)
- Dasgupta (2007)
- Dasgupta and Ng (2006)
- Dasgupta (2005)

### **3.2.6 Manipuri**

- T. D. Singh and Bandyopadhyay (2008)

### **3.2.7 Gujarati**

- Patel and Gali (2008)

### **3.2.8 Telugu**

- Karthik Kumar, Sudheer, and Avinesh (2006)
- Rama Sree, Uma Maheswara Rao, and Madhu Murthy (2008)

### **3.2.9 Burmese**

- Htay and Murthy (2008)
- Maung and Mikami (2008)

### **3.2.10 Uigur**

- Ablimit, M.Eli, and T.Kawahara (2008)

### **3.2.11 Hebrew**

- Bar-Haim, Sima'an, and Winter (2008)

### **3.2.12 Lao**

- Berment (2004)

### **3.2.13 Thai**

- Tongchim, Altmeyery, Sornlertlamvanich, and Isaharaz (2008)

### **3.2.14 Vietnamese**

- Nguyen et al. (2008)

### **3.2.15 Tagalog**

- Nelson (2004)

### **3.2.16 Turkmen**

- A. Cüneyd Tantuğ and Oflazer (2006)

### **3.2.17 Urdu**

- Humayoun, Hammarström, and Ranta (2007)
- Hardie (2003)
- Bögel, Butt, Hautli, and Sulger (2008)
- Hussain (2008)

### **3.2.18 Assamese**

- Sharma, Kalita, and Das (2002)

### **3.2.19 Oriya**

- Mohanty, Santi, and Das Adhikary (2005)

### **3.2.20 Sanskrit**

- Huet (2005)

### **3.2.21 Pashto**

- Khan and Zuhra (2007)

### **3.2.22 Syriac**

- Kiraz (1996)
- Kiraz (2001)
- Kiraz (2000)

### **3.2.23 Akkadian**

- Berthélemy (1998)
- Kataja and Koskeniemi (1988)

### **3.2.24 Mongolian**

- Khaltar and Fujii (2008)

### **3.2.25 Great Andamanese**

- Choudhary (2006)

## **3.3 Americas**

### **3.3.1 Greenlandic**

- Trosterud (2008b)

### **3.3.2 Iñupiaq**

- Trosterud (2008c)

### **3.3.3 Mapudungun**

- Monson et al. (2008)

### **3.3.4 Quiché**

- Kudlek (1975)

### **3.3.5 Ralámuri**

- Medina-Urrea (2006)

### **3.3.6 Chuj**

- Medina Urrea and Díaz (2003)

### **3.3.7 Aymara**

- Beesley (2003)

### **3.3.8 Cayuga**

- Graham (2007)

## **3.4 Africa**

### **3.4.1 Afrikaans**

- Stadler and Coetzer (1990)

### **3.4.2 Xhosa**

- Bosch, Pretorius, Podile, and Fleisch (2008)
- Bosch, Podile, Jones, and Mfusi (2003)

### **3.4.3 Swahili**

- Hurskainen (1992)
- Pauw, Schryver, and Wagacha (2006)

### **3.4.4 Sesotho**

- Johnson (2008)
- Schryver (2007)

### **3.4.5 Zulu**

- Bosch et al. (2008)
- Pretorius and Bosch (2003)

### **3.4.6 Ha**

- Harjula (2005)

### **3.4.7 Somali**

- Abdillahi, Nocera, and Torres-Moreno (2006)

### **3.4.8 Kikuyu**

- Pauw and Wagacha (2007)

### **3.4.9 Kinyarwanda**

- Muhirwe and Trosterud (2008)

### **3.4.10 Luo**

- Pauw, Wagacha, and Abade (2007)

### **3.4.11 Bambara**

- Fleisch and Seidel (2006)

### **3.4.12 Ekegusii**

- Elwell (2006)

### **3.4.13 Malagasy**

- Dalrymple, Liakata, and Mackie (2006)

Europe	22
Asia	25
Africa	13
Americas	8
	<hr/>
	68

Table 3: Number of low-density languages for which there is work on computational morphology divided up by continent.

## 4 Discussion

Table 3 shows a summary of the number of languages found in the survey.

There are about 70 low-density languages for which we have found work on computational morphological analysis. Undoubtedly, there is more work on which nothing has been published.

By far the least powerful language that has enjoyed the attention of a computational morphology implementation is Great Andamanese.

Among the 50 languages, it seems like the languages of Europe are over-represented in relation to their GLP. Surprisingly, there is no Australian Aboriginal language represented in spite of their presence in a high-technological country where there is government support and revitalization efforts for some (Larkin, 2005).

## 5 Conclusion

We have made a listing of work on computational morphology for low-density languages (defined as the economic power of its speakers). We hope this may be of use to better our understanding of the situation of the low-density languages in the technological age.

## References

- Abdillahi, N., Nocera, P., & Torres-Moreno, J.-M. (2006). Boites à outils tal pour les langues peu informatisées: le cas du somali. In *Journées d'analyses des données textuelles (jadt 06)* (p. 697-705). Besançon-France.

- Ablimit, M., M.Eli, & T.Kawahara. (2008). Partly supervised uighur morpheme segmentation. In *Proceedings of the oriental-cocosda workshop* (p. 71-76). Japan.
- A. Cüneyd Tantug, E. A., & Oflazer, K. (2006). Computer analysis of the turkmen language morphology. In T. Salakoski, F. Ginter, S. Pyysalo, & T. Pahikkala (Eds.), *Advances in natural language processing: Proceedings of the 5th international conference, fintal 2006 turku, finland, august 23-25, 2006* (Vol. 4139, p. 186-193). Springer-Verlag, Berlin.
- Alegria, I., Artola, X., Sarasola, K., & Urkia, M. (1996). Automatic morphological analysis of basque. *Literary & Linguistic Computing*, 11(4), 193-203.
- Anandan, P., Parthasarathi, K. S. R., & Geetha, T. V. (2002). Morphological analyzer for tamil. In *International conference on natural language processing, icon-2002, mumbai, december 18-21, 2002* (p. 3-10). Vikas Publishing House Pvt Ltd., New Delhi.
- Baldwin, T., Bird, S. G., & Hughes, B. (2006). Collecting low-density language materials on the web. In *Proceedings of 12th australasian web conference (ausweb06)*. Southern Cross University.
- Bar-Haim, R., Sima'an, K., & Winter, Y. (2008). Part-of-speech tagging of modern hebrew text. *Journal of Natural Language Engineering*, 14(2), 223-251.
- Beesley, K. R. (2003). *Finite-state morphological analysis and generation for aymara*. EACL Workshop on Finite State Methods in Natural Language Processing, Budapest, 2003.
- Beesley, K. R. (2004). Morphological analysis and generation: A first-step in natural language processing. In *First steps in language documentation for minority languages: Computational linguistic tools for morphology, lexicon and corpus compilation, proceedings of the saltmil workshop at lrec 2004* (p. 1-8). Lisboa, Portugal.
- Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues «peu dotées»*. Unpublished doctoral dissertation, Université Joseph-Fourier, Grenoble I.
- Berthélemy, F. (1998). A morphological analyzer for akkadian verbal forms with a model of phonetic transformations. In M. Rosner (Ed.), *Proceedings of the workshop on computational approaches to semitic languages 16th august 1998*. Université de Montreal, Montreal, Quebec, Canada.
- Bögel, T., Butt, M., Hautli, A., & Sulger, S. (2008). Developing a finite-state morphological analyzer for urdu and hindi. In *Proceedings of the sixth*

- international workshop on finite-state methods and natural language processing (fsmnlp 2007)*. Potsdam.
- Bosch, S., Podile, K., Jones, J., & Mfusi, M. (2003). Automating xhosa morphology. In *12th biennial international alasa (african language association of southern africa) conference, university of Stellenbosch, 8 July 2003*.
- Bosch, S., Pretorius, L., Podile, K., & Fleisch, A. (2008). Experimental fast-tracking of morphological analysers for Nguni languages. In *Proceedings of the sixth international language resources and evaluation (Irec'08)*. Marrakech, Morocco.
- Central Intelligence Agency of the United States. (2007). *The world factbook*. U.S. government.
- Choudhary, N. K. (2006). *Developing a computational framework for the verb morphology of Great Andamanese*. Unpublished master's thesis, Centre for Linguistics School of Language, Literature & Culture Studies Jawaharlal Nehru University New Delhi, India.
- Chrupała, G. (2008). *Towards a machine-learning architecture for lexical functional grammar parsing*. Unpublished doctoral dissertation, Dublin City University.
- Dalrymple, M., Liakata, M., & Mackie, L. (2006). Tokenization and morphological analysis for Malagasy. *Computational Linguistics and Chinese Language Processing*, 11(4), 315-332.
- Dandapat, S., Sarkar, S., & Basu, A. (2007, June). Automatic part-of-speech tagging for Bengali: An approach for morphologically rich languages in a poor resource scenario. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics companion volume proceedings of the demo and poster sessions* (pp. 221-224). Prague, Czech Republic: Association for Computational Linguistics.
- Dasgupta, S. (2005). Morphological analysis of inflectional compound words in Bangla. In *Ijcnlp'05*.
- Dasgupta, S. (2007). *Toward language-independent morphological segmentation and part-of-speech induction*. Unpublished master's thesis, The University of Texas at Dallas.
- Dasgupta, S., & Ng, V. (2006). Unsupervised morphological parsing of Bengali. *Language Resources and Evaluation*, 3-4, 311-330.
- Dasgupta, S., & Ng, V. (2007). Unsupervised word segmentation for Bangla. In *Proceedings of the 5th international conference on natural language processing (ICON 2007)*. Hyderabad, India.

- David, A., & Maxwell, M. (2008, January). Invited talk: Building language resources: Ways to move forward. In *Proceedings of the ijcnlp-08 workshop on nlp for less privileged languages* (pp. 1–2). Hyderabad, India: Asian Federation of Natural Language Processing.
- Dhonnchadha, E. U., Pháidín, C. N., & Genabith, J. V. (2003). Design, implementation and evaluation of an inflectional morphology finite state transducer for irish. *Machine Translation*, 18(3), 173-193.
- Elwell, R. (2006). Finite state methods for bantu verb morphology. In *Proceedings texas linguistics society x*.
- Erjavec, T., & Džeroski, S. (2004). Machine learning of morphosyntactic structure: Lemmatizing slovene words. *Applied Artificial Intelligence*, 18, 17-41.
- Fleisch, A., & Seidel, F. (2006). Cologne initiative on natural language processing in african languages. In *Proceedings of the lrec 2006 conference*. Genoa.
- Forsberg, M. (2007). *Three tools for language processing: Bnf converter, functional morphology, and extract*. Unpublished doctoral dissertation, Chalmers University of Technology, Gothenburg.
- Gordon, R. G., Jr. (Ed.). (2005). *Ethnologue: Languages of the world* (15 ed.). SIL International, Dallas.
- Graham, D. (2007). *Finite-state parsing of cayuga morphology*. Unpublished master's thesis, Memorial University of Newfoundland, Canada.
- Hajič, J. (2000). Morphological tagging: data vs. dictionaries. In *Proceedings of the first conference on north american chapter of the association for computational linguistics* (pp. 94–101). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Hardie, A. (2003). *The computational analysis of morphosyntactic categories in urdu*. Unpublished doctoral dissertation, University of Lancaster.
- Harjula, L. (2005). Morphological parsing of tone: An experiment with two-level morphology on the ha language. *Nordic Journal of African Studies*, 14(4), 452-463.
- Herath, S., Ikeda, T., Yokoyama, S., Isahara, H., & Ishizaki, S. (1989). Sinhalese morphological analysis: a step towards machine processing of sinhalese. In *Ieee international workshop on tools for artificial intelligence: Architectures, languages and algorithms*, (p. 100 - 107).
- Htay, H. H., & Murthy, K. N. (2008). Myanmar word segmentation using syllable level longest matching. In *The 6th workshop on asian language resources* (p. 41-48).

- Huet, G. (2005). A functional toolkit for morphological and phonological processing, application to a sanskrit tagger. *J. Funct. Program.*, 15(4), 573–614.
- Humayoun, M., Hammarström, H., & Ranta, A. (2007). Urdu morphology, orthography and lexicon extraction. In *Caasl-2: The second workshop on computational approaches to arabic script-based languages, july 21-22, 2007, lsa 2007 linguistic institute, stanford university*.
- Hurskainen, A. (1992). A two-level computer formalism for the analysis of bantu morphology: An application to swahili. *Nordic Journal of African Studies*, 1(1), 87-119.
- Hussain, S. (2008). Resources for urdu language processing. In *The 6th workshop on asian language resources* (p. 99-100).
- Idicula, S. M., & David, P. S. (2007). A morphological processor for malayalam language. *South Asia Research*, 27(2), 173-186.
- Johnson, M. (2008, June). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the tenth meeting of acl special interest group on computational morphology and phonology* (pp. 20–27). Columbus, Ohio: Association for Computational Linguistics.
- Kaalep, H.-J., & Vaino, T. (2001). Complete morphological analysis in the linguists toolbox. In *Congressus nonus internationalis fenno-ugristarum pars v* (p. 9-16). Tartu.
- Karthik Kumar, G., Sudheer, K., & Avinesh, P. V. S. (2006). Comparative study of various machine learning methods for telugu part of speech tagging. In *Proceedings of the nlpai contest workshop during nwai-06, sigai, mumbai*.
- Kataja, L., & Koskenniemi, K. (1988). Finite-state description of semitic morphology: A case study of ancient akkadian. In *Coling '88* (p. 313-315). ACL.
- Khaltar, B.-O., & Fujii, A. (2008, January). A lemmatization method for modern mongolian and its application to information retrieval. In *Proceedings of the third international joint conference on natural language processing (ijcnlp 2008)* (p. 1-8). Hyderabad, India: Asian Federation of Natural Language Processing.
- Khan, M. A., & Zuhra, F. T. (2007). The computational morphology of pashto nouns. *South Asian Language Review*, XVII(1).
- Kiraz, G. A. (1996). Syriac morphology: From a linguistic model to a computational implementation. In R. Lavenant (Ed.), *Vii symposium syriacum 1996*. Orientalia Christiana Analecta, Rome.

- Kiraz, G. A. (2000). Multitiered nonlinear morphology using multitape finite automata: a case study on Syriac and Arabic. *Computational Linguistics*, 26(1), 77-105.
- Kiraz, G. A. (2001). *Computational nonlinear morphology: With emphasis on semitic languages*. Cambridge University Press.
- Krstev, C., Vitas, D., & Erjavec, T. (2004). Morpho-syntactic descriptions in multext-east - the case of serbian. In *Informatika 28* (p. 431-436). The Slovene Society Informatika, Ljubljana.
- Kudlek, M. (1975). Computer programs for generating and analyzing quiché verb phrases. In E. Cerulli & G. Della Ragione (Eds.), *Linguistica – folklore – storia americana – sociologia* (Vol. 3, p. 45-54). Tilgher, Genoa.
- Larkin, S. (2005). *National indigenous languages survey report* (Tech. Rep.). Australian Institute of Aboriginal and Torres Strait Islander Studies.
- Lee, J. (2008, August). A nearest-neighbor approach to the automatic analysis of ancient greek morphology. In *Conll 2008: Proceedings of the twelfth conference on computational natural language learning* (pp. 127–134). Manchester, England: Coling 2008 Organizing Committee.
- Loftsson, H. (2008a). Tagging icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, 31(1), 47-72.
- Loftsson, H. (2008b). Tagging icelandic text: An experiment with integrations and combinations of taggers. *Nordic Journal of Linguistics*, 40(2), 175-181.
- Maung, Z. M., & Mikami, Y. (2008, January). A rule-based syllable segmentation of myanmar text. In *Proceedings of the ijcnlp-08 workshop on nlp for less privileged languages* (pp. 51–58). Hyderabad, India: Asian Federation of Natural Language Processing.
- Medina-Urrea, A. (2006). Affix discovery by means of corpora: Experiments for spanish, czech, rálámuli and chuj. In A. Mehler & R. Köhler (Eds.), *Aspects of automatic text analysis* (Vol. 209, p. 277-299). Springer, Berlin.
- Medina Urrea, A., & Díaz, E. C. B. (2003). Características cuantitativas de la flexión verbal del chuj. *Estudios de Lingüística Aplicada*, 38, 15-31.
- Mohanty, S., Santi, P. K., & Das Adhikary, K. P. (2005). Analysis and design of oriya morphological analyser: Some tests with orinet. In *Proceedings of symposium on indian morphology, phonology and language engineering*. IIT Kharagpur, India.
- Monson, C., Llitjós, A. F., Ambati, V., Levin, L., Lavie, A., Alvarez, A.,

- et al. (2008). Linguistic structure and bilingual informants help induce machine translation of lesser-resourced languages. In *Proceedings of the sixth international language resources and evaluation (lrec'08)*. Marrakech, Morocco.
- Müürisepp, K., Puolakainen, T., Muischnek, K., Koit, M., Roosmaa, T., & Uibo, H. (2003). A new language for constraint grammar: Estonian. In *Recent advances in natural language processing* (p. 304-310). Borovets, Bulgaria.
- Muhirwe, J., & Trosterud, T. (2008, January). Finite state solutions for reduplication in kinyarwanda language. In *Proceedings of the ijcnlp-08 workshop on nlp for less privileged languages* (pp. 73–80). Hyderabad, India: Asian Federation of Natural Language Processing.
- Nelson, H. J. (2004). *A two-level engine for tagalog morphology and a structured xml output for pc-kimmo*. Unpublished master's thesis, Brigham Young University.
- Nguyen, T. M. H., Rossignol, M., Le, H. P., Dinh, Q. T., Vu, X. L., & Nguyen, C. T. (2008). Word segmentation of vietnamese texts: a comparison of approaches. In *Proceedings of the sixth international language resources and evaluation (lrec'08)*. Marrakech, Morocco.
- Paikens, P. (2007). Lexicon-based morphological analysis of latvian language. In *Proceedings of the 3rd baltic conference on human language technologies, kaunas, october, 2007*.
- Patel, C., & Gali, K. (2008, January). Part-of-speech tagging for gujarati using conditional random fields. In *Proceedings of the ijcnlp-08 workshop on nlp for less privileged languages* (pp. 117–122). Hyderabad, India: Asian Federation of Natural Language Processing.
- Pauw, G. D., Schryver, G.-M. de, & Wagacha, P. (2006). Data-driven part-of-speech tagging of kiswahili. In *Proceedings of text, speech and dialogue, 9th international conference* (Vol. 4188, p. 197-204). Springer-Verlag, Berlin.
- Pauw, G. D., & Wagacha, P. (2007). Bootstrapping morphological analysis of gĩkũyũ using maximum entropy learning. In *Proceedings of the eighth interspeech conference*. Antwerp, Belgium.
- Pauw, G. D., Wagacha, P., & Abade, D. (2007). Unsupervised induction of dholuo word classes using maximum entropy learning. In *Proceedings of the first international computer science and ict conference (coscit 2007)*. Nairobi, Kenya: University of Nairobi.
- Pretorius, L., & Bosch, S. E. (2003). Finite-state computational morphology:

- An analyzer prototype for zulu. *Machine Translation*, 8(3), 195-216.
- Prószéky, G., & Novák, A. (2005). Computational morphologies for small uralic languages. In A. Arppe et al. (Eds.), *Inquiries into words, constraints and contexts: Festschrift for kimmo koskenniemi on his 60th birthday* (p. 116-125). CSLI Publications.
- Rama Sree, R., Uma Maheswara Rao, G., & Madhu Murthy, K. V. (2008). Assessment and development of pos tag set for telugu. In *The 6th workshop on asian language resources* (p. 85-88).
- Rimkutė, E., Daudaravičius, V., & Utkā, A. (2007, June). Morphological annotation of the lithuanian corpus. In *Proceedings of the workshop on balto-slavonic natural language processing* (pp. 94-99). Prague, Czech Republic: Association for Computational Linguistics.
- Sarkar, S., & Bandyopadhyay, S. (2008, January). Design of a rule-based stemmer for natural language text in bengali. In *Proceedings of the ijcnlp-08 workshop on nlp for less privileged languages* (pp. 65-72). Hyderabad, India: Asian Federation of Natural Language Processing.
- Schmid, H., & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Coling-08*. ACL.
- Schryver, G. D. P. G.-M. de. (2007). Dictionary writing system (dws) + corpus query package (cqp): The case of tshwanelex. *Lexikos*, 17, 226-246.
- Sharma, U., Kalita, J., & Das, R. (2002). Unsupervised learning of morphology for building lexicon for a highly inflectional language. In *Proceedings of the 6th workshop of the acl special interest group in computational phonology (siphon), philadelphia, july 2002* (p. 1-10). Association for Computational Linguistics.
- Simov, K., Osenova, P., Kolkovska, S., Balabanova, E., & Doikoff, D. (2004). A language resources infrastructure for bulgarian. In *Proceedings of lrec 2004* (p. 1685-1688.). Lisbon, Portugal.
- Singh, A. K. (2008, January). Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the ijcnlp-08 workshop on nlp for less privileged languages* (pp. 7-12). Hyderabad, India: Asian Federation of Natural Language Processing.
- Singh, T. D., & Bandyopadhyay, S. (2008, January). Morphology driven manipuri pos tagger. In *Proceedings of the ijcnlp-08 workshop on nlp for less privileged languages* (pp. 91-98). Hyderabad, India: Asian

- Federation of Natural Language Processing.
- Slavcheva, M. (2003). Some aspects of the morphological processing of bulgarian. In *Eacl 2003: Workshop morphological processing of slavic languages*.
- Sornlertlamvanich, V. (2008, January). Invited talk: Cross language resource sharing. In *Proceedings of the ijcnlp-08 workshop on nlp for less privileged languages* (pp. 3–4). Hyderabad, India: Asian Federation of Natural Language Processing.
- Stadler, L. G. de, & Coetzer, M. W. (1990). A morphological parser for afrikaans. In *Proceedings of the 13th conference on computational linguistics* (pp. 85–88). Morristown, NJ, USA: Association for Computational Linguistics.
- Streiter, O., Scannell, K. P., & Stuflessner, M. (2006). Implementing nlp projects for noncentral languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4), 267–289.
- Sulger, S. (2008). Implementing a finite-state morphological analyzer for irish: Issues at the morphology-syntax interface. In *Proceedings of the seventh international workshop on finite-state methods and natural language processing (fsmnlp 2008)*. Konstanz.
- Tongchim, S., Altmeyery, R., Sornlertlamvanich, V., & Isaharaz, H. (2008). A dependency parser for thai. In *Proceedings of the sixth international language resources and evaluation (lrec'08)*. Marrakech, Morocco.
- Trosterud, T. (2008a). *Faroese language technology*. <http://giellatekno.uit.no/fao.html> accessed 24 Dec 2008.
- Trosterud, T. (2008b). *Greenlandic language technology*. <http://giellatekno.uit.no/kal.html> accessed 24 Dec 2008.
- Trosterud, T. (2008c). *Iñupiaq language technology*. <http://giellatekno.uit.no/ipk.html> accessed 24 Dec 2008.
- Trosterud, T. (2008d). *Sámi language technology*. <http://giellatekno.uit.no/> accessed 24 Dec 2008.
- Uibo, H. (2002). Experimental two-level morphology of estonian. In *Proceedings of lrec 2002: Third international conference on language resources and evaluation* (p. 1012-1015). Las Palmas, Gran Canaria.
- Vikram, T. N., & Urs, S. R. (2007). Development of prototype morphological analyzer for the south indian language of kannada. In *Asian digital libraries. looking back 10 years and forging new frontiers* (Vol. 4822, p. 109-116). Springer-Verlag, Berlin.
- Viswanathan, S., Kumar, S. R., Shanmugam, B. K., & Arulmozi, S. (2003). A

tamil morphological analyser. In R. Sangal, S. Bendre, & U. N. Singh (Eds.), *Recent advances in natural language processing: Proceedings of the international conference natural icon-2003* (p. 31-39). Vikas Publishing House Pvt Ltd., New Delhi.

Yli-Jyrä, A. (2005). Toward a widely usable finite-state morphology workbench for less studied languages – part i: Desiderata. *Nordic Journal of African Studies*, 14(4), 479-491.