Supplement to ”On hapax legomena and morphological productivity”
Summary of Materials

Janet B. Pierrehumbert
Dept. of Engineering Science
University of Oxford
`janet.pierrehumbert@oerc.ox.ac.uk`

Ramon Granell
Dept. of Engineering Science
University of Oxford
`ramon.granell@oerc.ox.ac.uk`

1 Inclusion criteria and affix lists

The list of target morphemes was developed for a related (and unpublished) project on word embeddings in derivational morphology and compounding. The corpus for the study is the August 2013 dump of Wikipedia preprocessed as described in Levy and Goldberg (2014). Morphemes were selected by analyzing words in that occur at least 100 times in order to find all prefixes and suffixes that meet objective statistical criteria, with minimal dependence on data curation. The criteria are conservative, omitting many affixes that are widely held to be part of English morphology. They were designed to ensure that the affix is at least somewhat productive, and to minimize the number of spurious morphological parses using uniform and objective criteria (rather than intuitive judgments of individual words).

In the whole corpus, words are considered to be Real words if they occur at least twice in Wikipedia and contain only alphabetic characters, with the possible addition of one hyphen but no further punctuation marks. If there is variation in the use of a hyphen, only the more frequent variant is retained. There are 2,609,189 Real words representing 1.24 billion word tokens. A Real word is considered as potentially parseable into Prefix+Stem or Stem+Suffix if it contains at least six characters (or seven with at least three characters on either side of the hyphen), and the parse meets the following criteria:

- The putative affix and Stem both have at least three characters.
- The Stem is one of the Real words.
- The Stem has a higher frequency than the potentially complex word.

This frequency criterion is imposed because when it is violated, the parse is often opaque or spurious (Hay, 2001). The criterion eliminates parses such as *season* = *sea+son*. In the case of suffixes that begin in a vowel, we also allowed for the possibility of e-elision, as in *acclimatized* = *acclimate+ized*. If this alternation yielded multiple candidates for the Stem, we took the more frequent one. We did not make adjustments for consonant doubling, as in *stop, stopping*.

To select the prefixes and suffixes, we now look at the list of all words occurring at least 100 times (eg. in words with embeddings), with or without a valid parse. We impose the following criteria in order to identify affixes that are reliably parsable and to minimise statistical dependencies amongst the affixes in the study.

- The affix appears in at least 50 parsed word forms on the list.
- The affix is more likely to appear (as a character string) in a parsed word form than in a form that was not parsed, as determined by a t-test with a significance threshold of $P < 0.01$; its appearance reliably indicates that morpheme boundary is more likely than not to be present.
- If two affixes are spelling variants, we took only the more frequent one.
- If one suffix is an inflectional variant of another, we took the more frequent one.
- We also took the more frequent affix in in affix pairs such as *+stone, +tone* that could create ambiguous decompositions (*moon+stone*, not *moons+tone*).

These criteria result in 68 prefixes and 65 suffixes. Descriptive statistics for each are listed in the following tables.

Suffix	#Raw	#Real	#Type_All	#Type_Tail	#Tok_All	#Tok_Tail
able	15144	5331	2042	1258	1025923	5121
american	1257	356	335	194	77208	797
back	2042	728	589	364	154672	1493
based	14705	6312	6169	4679	139574	18615
board	1384	492	399	247	132750	1059
born	3788	1255	1146	817	51632	3442
bridge	1460	652	537	323	33432	1403
bury	1471	787	574	324	54541	1362
dale	2467	1395	1079	670	56491	3127
day	3927	937	547	370	234121	1476
don	5567	2454	1190	763	68740	3216
down	1593	653	567	368	98183	1534
ers	46029	18813	6881	4209	3152105	17992
ette	4149	2052	1306	899	67836	3684
field	3813	1765	1395	768	222639	3381
fish	1096	519	453	246	35830	1074
ford	3778	1691	1175	639	154927	2817
ful	2046	745	430	263	486341	1002
head	2596	1037	891	566	83110	2399
hill	2323	1057	862	553	37654	2307
house	2611	984	850	527	141036	2200
ingly	1481	688	477	292	129039	1234
ings	7678	3191	1750	1134	612613	4621
ington	2393	1138	707	317	106407	1458
ism	11225	4934	2594	1659	414085	6953
ist	16916	6657	2424	1604	968298	6474

Table 1: Statistics about suffixes (page 1 of 2). **#Raw** is the number of different word types (as defined by the presence of white space) in the original Wikipedia corpus that contain the suffix character sequence, including forms that only occur once and forms with internal numbers and punctuation marks. **#Real** is the number of words containing the suffix character sequence that we consider to be Real words: They occur at least twice, and are comprised of alphabetic characters with at most one hyphen. **#Type_All** is the number of words from #Real meeting all other inclusion criteria to be parsed as complex words containing the suffix, as described above. **#Type_Tail** is the number of words from #Type_All with frequencies in the range [2,11] **#Tok_All** is the number of tokens (additions of frequencies) that correspond to the words in #Type_All. **#Tok_Tail** is the number of tokens for the words in #Type_Tail.

Suffix	#Raw	#Real	#Type_All	#Type_Tail	#Tok_All	#Tok_Tail
istic	2304	916	431	284	164566	1136
ization	3757	1652	882	527	165568	2174
ized	7719	3070	1007	637	345352	2495
kin	4562	2234	1278	866	56778	3599
land	9986	4233	2679	1730	476841	7344
less	4568	1655	1412	874	246821	3729
ley	8073	3842	2365	1240	278552	5562
like	15628	5673	5518	4337	73467	17109
line	7502	2389	1330	868	270654	3598
man	22274	9992	6339	3869	751993	17069
mann	4363	2470	1818	1198	63806	5397
more	2768	1017	759	485	104376	2012
ness	9446	4014	2561	1574	375477	6586
net	6549	2994	1692	1230	54704	5044
off	4955	2474	1462	1037	98136	4397
out	5472	1680	1068	747	897660	2948
point	1619	556	501	323	53285	1428
port	3887	1351	752	481	252497	1842
related	8635	3590	3462	2774	46602	11013
sburg	1389	726	548	304	66724	1344
sey	1782	853	572	343	39886	1501
ship	3744	1364	821	530	703888	2234
shire	1045	405	246	150	168320	543
side	3239	1054	797	546	446985	2289
son	17727	7087	3593	2279	565715	9445
star	2051	908	726	473	65094	1977
stein	2923	1487	992	646	42831	2826
ston	3465	1577	921	550	107838	2432
stone	2413	1136	902	591	104678	2506
style	8661	3056	2942	2523	70014	9217
time	2434	727	579	380	280344	1426
town	4258	2064	1755	1130	205543	4861
ville	9190	5018	3993	2541	217306	11196
water	1415	471	399	237	86967	1011
way	4144	1597	1006	628	624389	2526
well	2946	1204	879	495	92237	2184
wood	3556	1785	1446	849	143385	3846
work	2096	646	454	302	102835	1218
worth	1463	744	552	277	46991	1196
Total	376947	156334	98808	64908	17698292	270501

Table 2: Statistics about suffixes (page 2 of 2). **#Raw** is the number of different word types (as defined by the presence of white space) in the original Wikipedia corpus that contain the suffix character sequence, including forms that only occur once and forms with internal numbers and punctuation marks. **#Real** is the number of words containing the suffix character sequence that we consider to be Real words: They occur at least twice, and are comprised of alphabetic characters with at most one hyphen. **#Type_All** is the number of words from #Real meeting all other inclusion criteria to be parsed as complex words containing the suffix, as described above. **#Type_Tail** is the number of words from #Type_All with frequencies in the range [2,11] **#Tok_All** is the number of tokens (additions of frequencies) that correspond to the words in #Type_All. **#Tok_Tail** is the number of tokens for the words in #Type_Tail.

Prefix	#Raw	#Real	#Type_All	#Type_Tail	#Tok_All	#Tok_Tail
air	4671	1525	1112	747	259208	3028
anti	15354	5915	4933	3717	191110	14746
ash	3845	1473	741	523	62762	2277
auto	6103	2320	1727	1292	106267	5055
back	3228	1058	797	520	186951	2252
bio	5977	2261	1692	1178	95990	4880
black	3273	1167	974	628	111658	2615
blue	2584	904	700	481	40380	2037
counter	2638	914	877	635	73646	2433
cross	3599	1488	1327	904	77752	3763
double	2981	1303	1230	940	29010	3724
down	1876	583	441	276	182217	1093
fire	2438	838	736	484	100494	2063
five	1705	653	621	393	36064	1586
fore	2533	810	410	231	99611	982
four	3403	1391	1024	661	72947	2806
free	3732	1269	871	569	109871	2365
gold	2897	1128	718	504	48082	2053
green	2820	1043	811	522	71176	2144
half	4300	1672	1618	1213	72914	4665
hand	3889	1393	996	693	114883	2744
hard	2769	948	591	391	63636	1616
head	2306	652	523	336	85109	1342
high	5626	2254	1381	936	376632	3791
home	2405	769	573	407	98834	1690
inter	10944	3842	2310	1523	1220232	6244
land	4508	1723	911	631	118065	2562
long	4537	1845	1257	842	139543	3383
low	3882	1573	1005	683	80082	2857
micro	6484	2673	2111	1519	88180	6069
mid	6694	2275	1341	917	189482	3765
mis	8956	3155	1613	1080	230526	4379
multi	7100	2748	2092	1442	112686	5812
news	2328	662	499	335	111918	1373
non	31413	12065	11360	8498	414731	34474
north	2537	742	428	265	275215	1070

Table 3: Statistics about prefixes (page 1 of 2). **#Raw** is the number of different word types (as defined by the presence of white space) in the original Wikipedia corpus that begin with the prefix character sequence, including forms that only occur once and forms with internal numbers and punctuation marks. **#Real** is the number of words containing the prefix character sequence that we consider to be Real words: They occur at least twice, and are comprised of alphabetic characters with at most one hyphen. **#Type_All** is the number of words from #Real meeting all other inclusion criteria to be parsed as complex words containing the prefix, as described above. **#Type_Tail** is the number of words from #Type_All with frequencies in the range [2,11] **#Tok_All** is the number of tokens (additions of frequencies) that correspond to the words in #Type_All. **#Tok_Tail** is the number of tokens for the words in #Type_Tail.

Prefix	#Raw	#Real	#Type_All	#Type_Tail	#Tok_All	#Tok_Tail
off	4281	1214	586	399	106646	1591
one	4563	1272	1047	694	108602	2960
out	4944	1344	919	541	703583	2200
over	8089	2714	2334	1557	518395	6278
photo	3045	1116	878	610	46257	2436
post	9739	3333	2854	2197	125956	8622
red	6355	2413	1287	868	81266	3493
sand	3773	1524	904	635	63908	2579
sea	6406	2072	1079	720	131239	2952
second	1317	429	365	230	36235	957
self	5674	2835	2741	1897	171397	8103
semi	6521	2602	2299	1808	107159	6937
short	1712	641	506	323	90707	1280
side	2049	748	617	435	44488	1750
single	2697	1267	1218	858	51215	3583
six	2361	918	611	390	37361	1789
sky	2238	889	705	458	34779	2061
south	2278	669	400	245	253277	1024
star	5121	1782	1094	736	192489	2996
sub	14260	5075	3520	2526	423130	10076
sun	5719	2200	1254	856	170871	3554
super	8580	3156	2633	1903	210169	7681
three	2874	1128	1082	685	86698	2876
time	3942	969	768	567	54096	2365
two	3617	1376	1257	814	121388	3428
under	4328	1434	1241	837	418966	3384
water	3247	1165	1005	709	117039	2860
well	3059	1319	1104	678	143513	2754
west	3969	1487	796	519	81102	2111
white	2326	861	751	465	50829	2036
wiki	8466	2281	1608	1262	134623	4830
wood	2651	1014	719	468	93488	1979
Total	336536	122281	92533	64806	10758735	263263

Table 4: Statistics about prefixes (page 2 of 2). **#Raw** is the number of different word types (as defined by the presence of white space) in the original Wikipedia corpus that begin in the prefix character sequence, including forms that only occur once and forms with internal numbers and punctuation marks. **#Real** is the number of words containing the prefix character sequence that we consider to be Real words: They occur at least twice, and are comprised of alphabetic characters with at most one hyphen. **#Type_All** is the number of words from #Real meeting all other inclusion criteria to be parsed as complex words containing the prefix, as described above. **#Type_Tail** is the number of words from #Type_All with frequencies in the range [2,11] **#Tok_All** is the number of tokens (additions of frequencies) that correspond to the words in #Type_All. **#Tok_Tail** is the number of tokens for the words in #Type_Tail.

2 Data files

The word lists are in .csv format. The list for the prefixed words is in vocWithWordFreqprefix.csv. Column A is the whole word; Column B is the prefix; Column C is the stem; and Column D is the frequency for the whole word (as a count).

The list for the suffixed words is in vocWithWordFreqsuffix.csv. Column A is the whole word; Column B is the stem; Column C is the suffix; and Column D is the frequency for the whole word (as a count).

References

Jennifer Hay. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics*, 39(6; ISSU 376):1041–1070.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.