Estimating Lexical Priors for Low-Frequency Morphologically Ambiguous Forms

Harald Baayen*
Max Planck Institute for Psycholinguistics

Richard Sproat†
Bell Laboratories

Given a form that is previously unseen in a sufficiently large training corpus, and that is morphologically n*-ways ambiguous (serves* n *different lexical functions) what is the best estimator for the lexical prior probabilities for the various functions of the form? We argue that the best estimator is provided by computing the relative frequencies of the various functions among the* hapax legomena*—the forms that occur exactly once in a corpus; in particular, a hapax-based estimator is better than one based on the proportion of the various functions among words of* all *frequency ranges. As we shall argue, this is because when one computes an* overall *measure, one is including high-frequency words, and high-frequency words tend to have idiosyncratic properties that are not at all representative of the much larger mass of (productively formed) low-frequency words. This result has potential importance for various kinds of applications requiring lexical disambiguation, including, in particular, stochastic taggers. This is especially true when some initial hand-tagging of a corpus is required: for predicting lexical priors for very low-frequency morphologically ambiguous types (most of which would not occur in any given corpus), one should concentrate on tagging a good representative sample of the hapax legomena, rather than extensively tagging words of all frequency ranges.*

1. Introduction

As a number of writers on morphology have noted (most recently and notably Beard [1995]), it is common to find that a particular affix or other morphological marker serves more than one function in a language. For example, in many morphologically complex languages it is often the case that several slots in a paradigm are filled with the same form; put in another way, it is common to find that a particular morphological form is in fact ambiguous between several distinct functions. This phenomenon—which in the domain of inflectional morphology is termed **syncretism**—can be illustrated by a Dutch example such as *lopen* 'walk', which can either be the infinitive form ('to walk') or the finite plural (present tense) form ('we, you, or they walk'). In some cases, syncretism is completely systematic: for example the case cited in Dutch, where the *-en* suffix can always function in the two ways cited; or in Latin, where the plural dative and ablative forms of nouns and adjectives are always identical, no matter what paradigm the noun belongs to. In other cases, a particular instance of syncretism may be displayed only in some paradigms: for example, Russian feminine nouns, such as *loshad'* 'horse' (Cyrillic лошадь), have the same form for both the genitive singular

* Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands. E-mail: baayen@mpi.nl
† 600 Mountain Avenue, Murray Hill, NJ 07974, USA. E-mail: rws@research.att.com

— *loshadi* (Cyrillic лошади) — and the nominative plural, whereas masculine nouns typically distinguish these forms. In still other cases, the syncretism may be partial in that two forms may be identical at one level of representation — say, orthography — but not another — say, pronunciation. For example the written form *goroda* in Russian (Cyrillic города) may either be the nominative plural or the genitive singular of 'city'. In the genitive singular, the stress is on the first syllable (/g'orədʌ/), whereas in the nominative plural the stress resides on the final syllable (/gərʌd'a/); note that the difference in stress results in very different vowel qualities for the two forms, as indicated in the phonetic transcriptions.

Syncretism and related morphological ambiguities present a problem for many NL applications where lexical disambiguation is important; cases where the orthographic form is identical but the pronunciations of the various functions differ are particularly important for speech applications, such as text-to-speech, since appropriate word pronunciations must be computed from orthographic forms that underspecify the necessary information. Ideally one would like to build models that use contextual information to perform lexical disambiguation (Yarowsky 1992, 1994), but such models must be trained on specialized tagged corpora (either hand-generated or semi-automatically generated) and such training corpora are often not available, at least in the early phases of constructing a particular application. Lacking good contextual models, one is forced to fall back on estimates of the **lexical prior** probabilities for the various functions of a form. Following standard terminology, a lexical prior can be defined as follows: Imagine that a given form is *n*-ways ambiguous; the lexical prior probability of sense *i* of this form is simply the probability of sense *i* independent of the context in which the particular instantiation of the form occurs. Assuming one has a tagged corpus, one can usually get reasonable estimates of the lexical priors for the frequent forms (such as Dutch *lopen* 'walk') by simply counting the number of times the form occurs in each of its various functions and dividing by the total number of instances of the form (in any function). This yields the **Maximum Likelihood Estimate** (MLE) for the lexical prior probability. But for infrequent or unseen forms, it is less clear how to compute the estimate. Consider another Dutch example like *aanlokken* 'entice, appeal'. This form occurs only once, as an infinitive, in the Uit den Boogaart (henceforth UdB) corpus (Uit den Boogaart 1975); in other words it is a **hapax legomenon** (< Greek *hapax* 'once', *legomenon* 'said') in this corpus. Obviously the lexical prior probability of this form expressing the finite plural is not zero, the MLE is a poor estimate in such cases. When one considers forms that do not occur in the training corpus (e.g., *bedraden* 'to wire') the situation is even worse. The problem, then, is to provide a more reasonable estimate of the relative probabilities of the various potential functions of such forms.[1]

2. Estimating the Lexical Priors for Rare Forms

For a common form such as *lopen* 'walk' a reasonable estimate of the lexical prior probabilities is the MLE, computed over all occurrences of this form. So, in the UdB corpus, *lopen* occurs 92 times as an infinitive and 43 times as a finite plural, so the MLE

[1] Even models of disambiguation that make use of context, such as statistical *n*-gram taggers, often presume some estimate of lexical priors, in addition to requiring estimates of the transition probabilities of sequences of lexical tags (Church 1988; DeRose 1988; Kupiec 1992), and this again brings up the question of what to do about unseen or low-frequency forms. In working taggers, a common approach is simply to apply a uniform small probability to the various senses of unseen or low-frequency forms: this was done in the tagger discussed in Church (1988), for example.

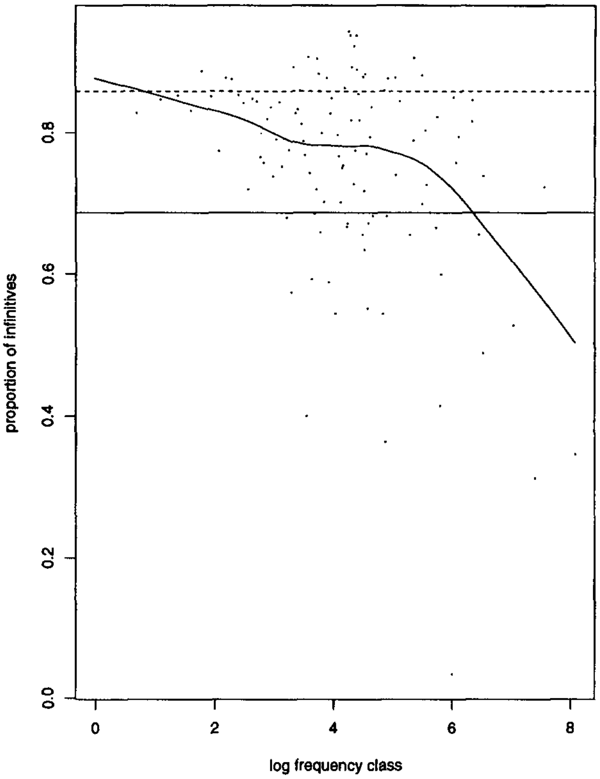

Figure 1
Relative frequency of Dutch infinitives versus finite plurals in the Uit den Boogaart corpus, as a function of the (natural) log of the frequency of the word forms. The horizontal solid line represents the overall MLE, the relative frequency of the infinitive as computed over all tokens; the horizontal dashed line represents the relative frequency of the infinitive among the hapax legomena. The solid curve represents a locally weighted regression smoothing (Cleveland 1979).

estimate of the probability of the infinitive is 0.68. For low-frequency forms such as *aanlokken* or *bedraden*, one might consider basing the MLE on the aggregate counts of all ambiguous forms in the corpus. In the UdB corpus, there are 21,703 infinitive tokens, and 9,922 finite plural tokens, so the MLE for *aanlokken* being an infinitive would be 0.69. Note, however, that the application of this overall MLE presupposes that the relative frequencies of the various functions of a particular form are independent of the frequency of the form itself. For the Dutch example at hand, this presupposition predicts that if we were to classify *-en* forms according to their frequency, and then for each frequency class thus defined, plot the relative frequency of infinitives and finite plurals, the regression line should have a slope of approximately zero.

2.1 Dutch Verb Forms in *-en*

Figure 1 shows that this prediction is not borne out. This scatterplot shows the relative frequency of the infinitive versus the finite plural, as a function of the log-frequency of the *-en* form. At the left-hand edge of the graph, the relative frequency of the infinitives for the hapax legomena is shown. This proportion is also highlighted by the dashed horizontal line. As we proceed to the right, we observe that there is a general downward curvature representing a lowering of the proportion of infinitives for the higher-frequency words. This trend is captured by the solid nonparametric regression line; an explanation for this trend will be forthcoming in Section 3. (It will be noted that in Figure 1 the variance is fairly small for the lower-frequency ranges, higher for the middle ranges, and then small again for the high-frequency ranges; anticipating somewhat, we note the same trends in Figures 2 and 3. This variance pattern follows from the high variability in the absolute numbers of types realized, especially in the middle log-frequency classes, in combination with the assumption that for any log-frequency class, the proportion for that class is itself a random variable.) The solid horizontal line represents the proportion of infinitives calculated over *all* frequency classes, and the dashed horizontal line represents the proportion of infinitives calculated over just the hapax legomena. The two horizontal lines can be interpreted as MLEs for the probability of an *-en* form being an infinitive: the solid line or **overall** MLE clearly provides an estimate based on the whole population, whereas the dashed line or **hapax-based** MLE provides an estimate for the hapaxes. The overall MLE computes a lower relative frequency for the infinitives, compared to the hapax-based MLE. The question, then, is: Which of these MLEs provides a better estimate for low-frequency types? In particular, for types that have not been seen in the training corpus, and for which we therefore have no direct estimate of the word-specific prior probabilities, we would like to know whether the hapax-based or overall MLE provides a better estimate.

To answer this question we compared the accuracy of the overall and hapax-based MLEs using tenfold cross-validation. We first randomized the list of *-en* tokens from the UdB corpus, then divided the randomized list into ten equal-sized parts. Each of the ten parts was held out as the test set, and the remaining nine-tenths was used as the training set over which the two MLE estimates were computed. The results are shown in Table 1. In this table, $N_0(inf)$ and $N_0(pl)$ represent the observed number of tokens of infinitives and plurals in the held-out portion of the data, representing types that had not been seen in the training data. The final four rows compare the estimates for these numbers of tokens given the overall MLE ($E_o[N_0(inf)]$ and $E_o[N_0(pl)]$), versus the hapax-based MLE ($E_h[N_0(inf)]$ and $E_h[N_0(pl)]$). For all ten runs, the hapax-based MLE is clearly a far better predictor than the overall MLE.[2]

2.2 English Verb Forms in *-ed*

The pattern that we have observed for the Dutch infinitive-plural ambiguity can be replicated for other cases of morphological ambiguity. Consider the case of English verbs ending in *-ed*, which are systematically ambiguous between being simple past tenses and past participles. The upper panel of Figure 2 shows the distribution of the relative frequencies of the two functions, plotted against the natural log of the frequency for the Brown corpus (Francis and Kucera 1982). (All lines, including the nonparametric regression line are interpretable as in Figure 1.) Results of a tenfold cross-validation are shown in Table 2. Clearly, in this case the magnitude of the difference between the overall MLE and the hapax-based MLE is smaller than in the previous example: indeed in cross validations 6, 8, and 9, the overall MLE is superior. Nonetheless, the hapax-based MLE remains a significantly better predictor overall.[3]

[2] A paired t-test on the ratios $N_0(inf)/N_0(pl)$ versus $E_o[N_0(inf)]/E_o[N_0(pl)]$ reveals a highly significant difference ($t_9 = 13.4, p < 0.001$); conversely a comparison of $N_0(inf)/N_0(pl)$ and $E_h[N_0(inf)]/E_h[N_0(pl)]$ reveals no difference ($t_9 = 0.96, p > 0.10$).

[3] A paired t-test on the ratios $N_0(vbn)/N_0(vbd)$ versus $E_o[N_0(vbn)]/E_o[N_0(vbd)]$ reveals a significant difference ($t_9 = 2.47, p < 0.05$); conversely a comparison of $N_0(vbn)/N_0(vbd)$ and $E_h[N_0(vbn)]/E_h[N_0(vbd)]$ reveals no difference ($t_9 = 0.48, p > 0.10$).

Table 1
Results of tenfold cross-validation for Dutch *-en* verb forms from the Uit den Boogaart corpus. Columns represent different cross-validation runs. *N(inf)* and *N(pl)* are the number of tokens of the infinitives and finite plurals, respectively, in the training set. $N_1(inf)$ and $N_1(pl)$ are the number of tokens of the infinitives and finite plurals, respectively, among the hapaxes in the training set. OMLE and HMLE are, respectively, the overall and hapax-based MLEs. $N_0(inf)$ and $N_0(pl)$ denote the number of tokens in the held-out portion that have *not* been observed in the training set. The expected numbers of tokens of infinitives and plurals for types unseen in the training set, using the overall MLE are denoted as $E_o[N_0(inf)]$ and $E_o[N_0(pl)]$; the corresponding estimates using the hapax-based MLE are denoted as $E_h[N_0(inf)]$ and $E_h[N_0(pl)]$.

Run	1	2	3	4	5	6	7	8	9	10
$N(inf)$	19,509	19,527	19,536	19,526	19,507	19,511	19,533	19,524	19,569	19,585
$N(pl)$	8,953	8,935	8,926	8,936	8,955	8,952	8,930	8,939	8,894	8,878
OMLE	0.685	0.686	0.686	0.686	0.685	0.685	0.686	0.686	0.688	0.688
$N_1(inf)$	1,075	1,086	1,066	1,068	1,092	1,091	1,098	1,066	1,094	1,079
$N_1(pl)$	185	184	180	182	179	185	184	178	179	180
HMLE	0.853	0.855	0.856	0.854	0.859	0.855	0.856	0.857	0.859	0.857
$N_0(inf)$	120	114	133	125	133	123	102	118	121	127
$N_0(pl)$	24	19	20	18	18	16	15	23	23	21
$E_o[N_0(inf)]$	99	91	105	98	103	95	80	97	99	102
$E_o[N_0(pl)]$	45	42	48	45	48	44	37	44	45	46
$E_h[N_0(inf)]$	123	114	131	122	130	119	100	121	124	127
$E_h[N_0(pl)]$	21	19	22	21	21	20	17	20	20	21

2.3 Dutch Words in *-en*: A More General Problem

In the two examples we have just considered, the hapax-based MLE, while being a better predictor of the a priori lexical probability for unseen cases than the overall MLE, does not actually yield a different prediction as to which function of a form is more likely. This does not hold generally, however, and the bottom panel of Figure 2 presents a case where the hapax-based MLE does yield a different prediction as to which function is more likely. In this plot we consider Dutch word forms from the UdB corpus ending in *-en*. As we have seen, Dutch *-en* is used as a verb marker: it marks the infinitive, present plural, and for strong verbs, also the past plural; it is also used as a marker of noun plurals. The case of noun plurals is somewhat different from the preceding two cases since it is not, strictly speaking, a case of morphological syncretism. However, it is a potential source of ambiguity in text analysis, since a low frequency form in *-en*, where one may not have seen the stem of the word, could potentially be either a noun or a verb. Also, systematic ambiguity exists among cases of noun-verb conversion: for example *fluiten* is either a noun meaning 'flutes' or a verb meaning 'to play the flute'; *spelden* means either 'pins' or 'to pin'; and *ploegen* means either 'ploughs' or 'to plough'. Results for a tenfold cross-validation for these data are shown in Table 3.[4] In this case, the overall MLE would lead one to predict that for an unseen form in *-en*, the verbal function would be more likely. Contrariwise, the hapax-based MLE predicts that the nominal function would be more likely. Again, it is the hapax-based MLE that proves to be superior.

[4] A paired t-test on the ratios $N_0(v)/N_0(n)$ versus $E_o[N_0(v)]/E_o[N_0(n)]$ reveals a highly significant difference ($t_9 = 95.95, p < 0.001$); conversely a comparison of $N_0(v)/N_0(n)$ and $E_h[N_0(v)]/E_h[N_0(n)]$ reveals no difference ($t_9 = 0.12, p > 0.10$).

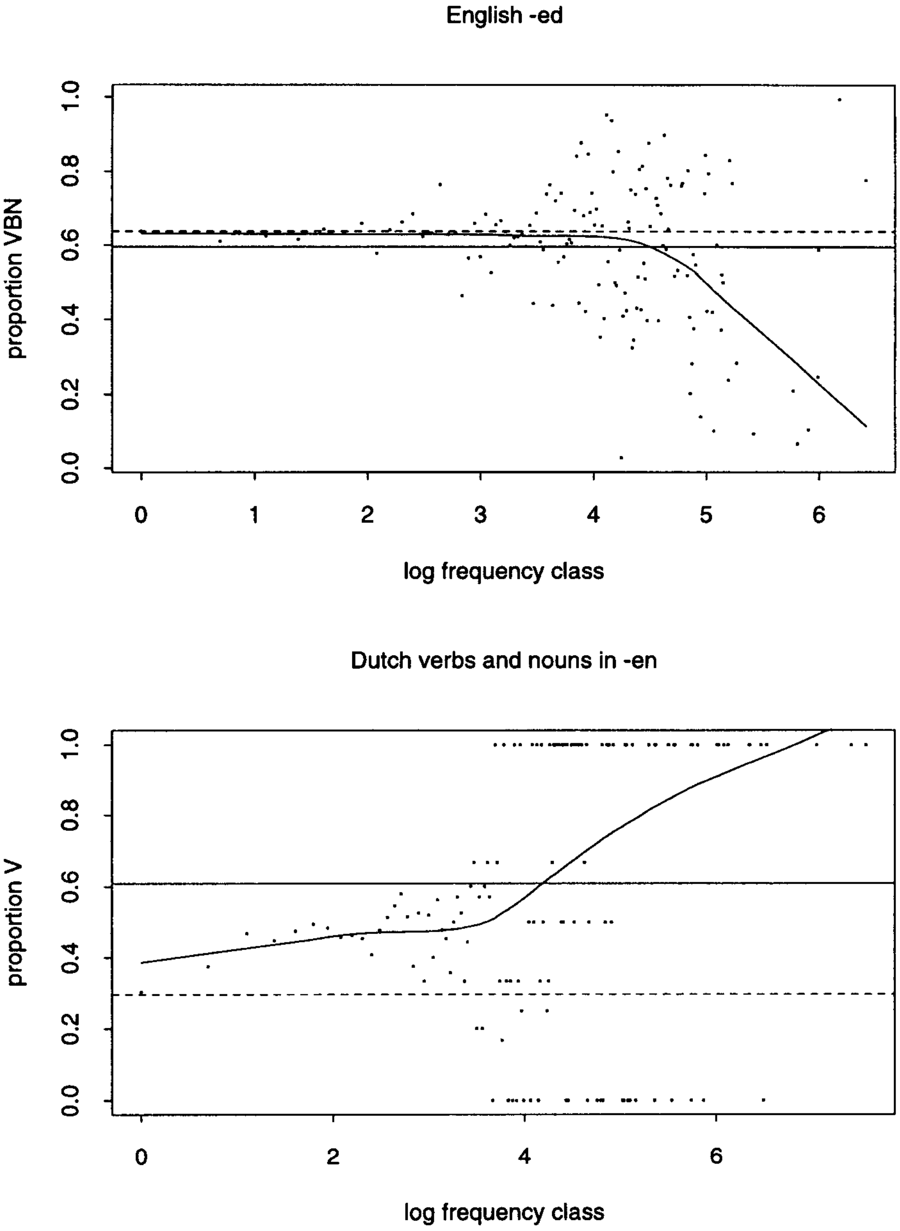

Figure 2
The top panel displays the distribution in the Brown corpus of the relative frequencies of English simple past tense verbs in *-ed* (Brown corpus tag *VBD*) versus past participles in *-ed* (*VBN*), plotted against log-frequency. The bottom panel displays the relative frequency as a function of log-frequency of Dutch verbs in *-en* (infinitives, present plurals, and strong past tense plurals), versus plural nouns in *-en*, computed over the Uit den Boogaart corpus. Lines are interpreted as in Figure 1.

Table 2
Cross-validation statistics for English past participles versus simple past tense verbs.

Run	1	2	3	4	5	6	7	8	9	10
N(vbn)	20,386	20,360	20,376	20,372	20,388	20,451	20,431	20,431	20,426	20,400
N(vbd)	13,845	13,871	13,855	13,859	13,843	13,781	13,801	13,801	13,806	13,832
OMLE	0.596	0.595	0.595	0.595	0.596	0.597	0.597	0.597	0.597	0.596
N$_1$(vbn)	701	695	678	700	693	705	690	692	710	711
N$_1$(vbd)	395	401	405	406	406	403	404	405	393	403
HMLE	0.640	0.634	0.626	0.633	0.631	0.636	0.631	0.631	0.644	0.638
N$_0$(vbn)	80	86	101	83	71	61	85	75	72	77
N$_0$(vbd)	49	52	37	41	43	45	41	50	48	42
E$_o$[N$_0$(vbn)]	77	82	82	74	68	63	75	75	72	71
E$_o$[N$_0$(vbd)]	52	56	56	50	46	43	51	50	48	48
E$_h$[N$_0$(vbn)]	83	88	86	78	72	67	79	79	77	76
E$_h$[N$_0$(vbd)]	46	50	52	46	42	39	47	46	43	43

Table 3
Cross-validation statistics for Dutch verbs in *-en* versus plural nouns in *-en*.

Run	1	2	3	4	5	6	7	8	9	10
N(v)	25,237	25,283	25,267	25,245	25,292	25,267	25,205	25,207	25,261	25,294
N(n)	18,306	18,260	18,277	18,299	18,252	18,277	18,339	18,337	18,283	18,250
OMLE	0.580	0.581	0.580	0.580	0.581	0.580	0.579	0.579	0.580	0.581
N$_1$(v)	1,312	1,295	1,287	1,317	1,284	1,298	1,298	1,297	1,292	1,298
N$_1$(n)	2,913	2,910	2,939	2,942	2,901	2,922	2,979	2,969	2,936	2,931
HMLE	0.311	0.308	0.305	0.309	0.307	0.308	0.303	0.304	0.306	0.307
N$_0$(v)	124	131	154	142	148	143	148	156	153	139
N$_0$(n)	325	344	327	334	352	335	289	301	327	319
E$_o$[N$_0$(v)]	260	276	279	276	290	277	253	265	278	266
E$_o$[N$_0$(n)]	189	199	202	200	210	201	184	192	202	192
E$_h$[N$_0$(v)]	139	146	146	147	153	147	133	139	147	141
E$_h$[N$_0$(n)]	310	329	335	329	347	331	304	318	333	317

2.4 Disyllabic Dutch Words Ending in *-er*

One final example—also not a case of syncretism—concerns the ambiguity of the sequence *-er* in Dutch, which occurs word-finally in monomorphemic nouns (*moeder*, 'mother'), adjectives (*donker*, 'dark'), and proper names (*Pieter*, 'Peter'), but which is also used as a suffix to form comparatives (*sneller*, 'faster') and "agentive" nouns (*schrijver*, 'writer'). Since monomorphemic nouns and adjectives in this class are mostly disyllabic, we will restrict our attention to the disyllabic instances of words ending in *-er*. Again we find that the hapax-based MLE is superior to the overall MLE for predicting to which of these five categories an unseen disyllabic word belongs.

Table 4 lists the overall MLE, the hapax-based MLE and the statistics on which these estimates are based; Figure 3 plots the corresponding proportions as a function of log-frequency. Table 4 also lists the results of tenfold cross-validation by specifying, for each category, its contribution to the X^2-statistic summing over the ten cross-validation runs. (A more condensed format was chosen for this table than for the previous tables, since here we are dealing with a fivefold ambiguity; the previous format would have resulted in a rather large table in the present case.) Clearly, predictions based on the

Table 4
Results of tenfold cross-validation for Dutch disyllabic *-er* words. N and N_1 are the number of tokens and number of hapax legomena in the Uit den Boogaart corpus for simplex and complex adjectives and nouns, and proper names. OMLE and HMLE are, respectively, the overall and hapax-based MLEs based on N and N_1. For each category, the columns headed by X^2(OMLE) and X^2(HMLE) list the summed contribution to the X^2-measures over ten cross-validation runs for the overall and hapax-based estimates.

String type	N	N_1	OMLE	HMLE	X^2(OMLE)	X^2(HMLE)
Simplex noun in *-er*	2,157	43	0.438	0.206	46.52	1.63
Derived noun in *-er*	581	51	0.118	0.244	32.02	1.94
Simplex adjective in *-er*	486	6	0.099	0.029	18.22	14.90
Derived adjective in *-er*	1,409	41	0.286	0.196	14.97	9.05
Proper name in *-er*	291	68	0.059	0.325	361.22	5.98
	4,924	209	1.000	1.000	462.97	33.50

hapax-based MLE are superior to those based on the overall MLE ($X^2_{(36)} = 462.97, p < .001$ for the overall MLE, $X^2_{(36)} = 33.50, p > .5$ for the hapax-based MLE). In particular, proper names in *-er* have a much higher probability of occurrence than the overall MLE would suggest. Of course, in orthography-based applications, one can rely to some extent on capitalization to indicate proper names, so one might want to eliminate those from consideration here on this basis. Removing the category of proper names from the analysis, a cross-validation test again reveals significantly better predictions for the hapax-based MLE ($X^2_{(27)} = 43.61, p = .023$) than for the overall MLE ($X^2_{(27)} = 120.03, p < .001$).

2.5 Summary

We have demonstrated with four separate examples that the hapax-based MLE is superior to the overall MLE in predicting the proportions, among unseen forms, of the various functions of morphologically ambiguous categories. Could an even better estimator be obtained by taking not only the proportion for the hapax legomena into account but also the proportions for other low-log-frequency classes? To answer this question, note that the scatterplot in the bottom panel of Figure 2 reveals a downward curvature at the very left-hand side: even for the lowest-log-frequency classes, the likelihood of a word being a verb decreases with decreasing log-frequency. This suggests that for this particular example the hapax legomena alone should be used to estimate the probability that an unseen word is a noun or verb, rather than the hapax legomena in combination with other low-frequency classes (the words occurring twice, three times, etc.). Interestingly, the top panel of Figure 2 does not reveal even a hint of a trend among the lowest-log-frequency classes, and in Figure 1 the observed proportions for log-frequency less than 2 also do not reveal a clear pattern. For Figure 3, clear trends for the lower-log-frequency classes seem to obtain in all cases except the plot showing the proportion of simplex adjectives. Taken jointly, these observations suggest informally that an MLE based on the hapax legomena will never be inferior to MLEs that take additional log-frequency classes at the lower end of the log-frequency range into account. At the same time, the example of Dutch verb and noun forms in *-en* suggests that the hapax-based MLE can be superior to such MLEs—in this particular case, inclusion of these lower-frequency classes would bring the adjusted MLE more in line with the overall MLE, resulting in a loss of accuracy. These considerations lead us to conclude that the hapax-based MLE is to be preferred to an adjusted MLE that includes other low-log-frequency classes.

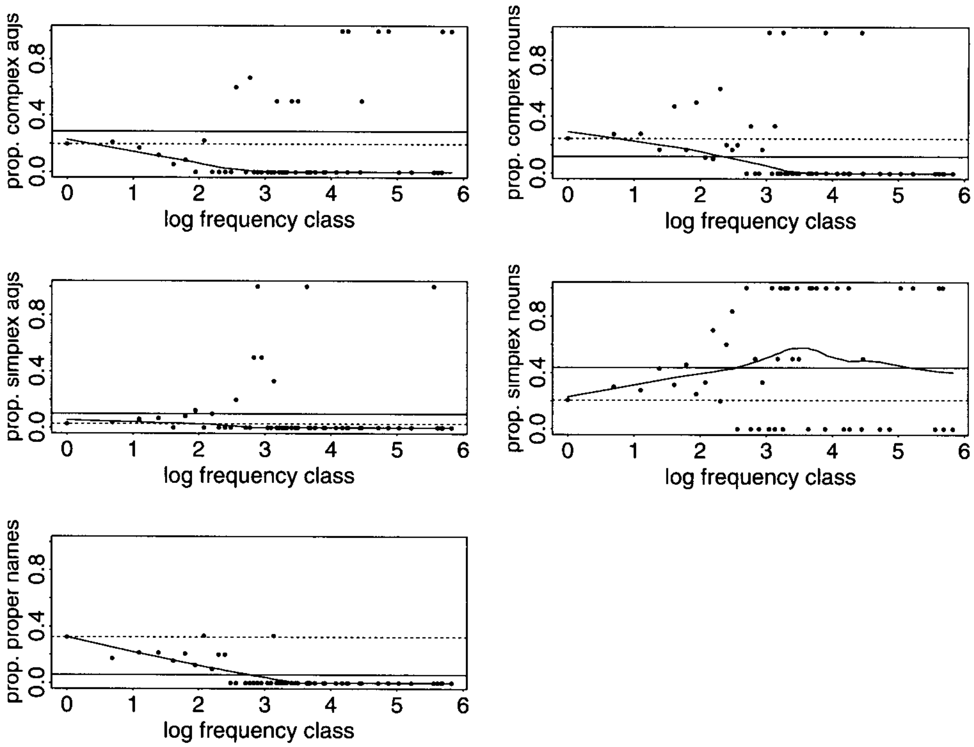

Figure 3
The distribution in the Uit den Boogaart corpus of the relative frequencies of disyllabic words in *-er*, plotted against log-frequency. Five types of words are distinguished: monomorphemic words in *-er* (*moeder*, 'mother'); bimorphemic nouns in *-er* (*schrijver*, 'writer'); monomorphemic adjectives in *-er* (*donker*, 'dark'); bimorphemic adjectives in *-er* (*groter*, 'greater'); and proper names in *-er* (*Pieter*, 'Peter'). Lines are interpreted as in Figure 1.

3. Discussion

As we have seen in the four examples discussed above, the MLE computed over hapax legomena yields a better prediction of lexical prior probabilities for unseen cases than does an MLE computed over the entire training corpus. We now have to consider why this result holds. As we shall see, the reasons are different from case to case, but nonetheless share a commonality: in all four cases, idiosyncratic lexical properties of high-frequency words dominate the statistical properties of the high-frequency ranges, thus making the overall MLE a less reliable predictor of the properties of the low-frequency and unseen cases.

First let us discuss the final case, that of *-er* ambiguity in Dutch, beginning with the derived and underived nouns. The hapax-based MLE estimate for derived nouns in *-er* is somewhat higher than the overall MLE; for underived nouns, the hapax-based MLE is significantly *lower* — half — of the overall MLE. This can be explained by the observation that a good many of the underived nouns in *-er* are high-frequency words such as *moeder* 'mother' and *vader* 'father'. Such words contribute to the overall proportional mass of the underived nouns, thus boosting the estimate of the overall MLE for this class. A similar argument holds for the derived and underived adjectives. Turning to proper names, we see that the hapax-based MLE is *much* larger than the overall MLE. Proper names differ from ordinary words in that there are relatively few proper names that are highly frequent, in comparison with words in general, but there are large numbers of types of names that occur rarely. Thus, we expect an imbalance of the kind we observe.

Consider next the ambiguity in Dutch between *-en* verb forms and *-en* plural nouns. Ceteris paribus, plural nouns are less frequent than singular nouns; on the other hand, *-en* for verbs serves both the function of marking plurality and of marking the infinitive. High-frequency verbs include some very common word forms, such as the auxiliaries *hebben* 'have', *zullen* 'will', *kunnen* 'can', and *moeten* 'must'. Thus, for the high-frequency ranges, the data is weighted heavily towards verbs. On the other hand, while both nouns and verbs are open classes, nouns are far more productive as a class than are verbs (Baayen and Lieber 1991), and this pattern becomes predominant in the low-frequency ranges: among low-frequency types, most tokens are nouns. Hence, for the low-frequency ranges, the data is weighted towards nouns. These two opposing forces conspire to yield a downward trend in the percentage of verbs as we proceed from the high- to the low-frequency ranges.

Next, consider the English past tense versus past participle ambiguity. One of the important functions of the past participle form is as an adjectival modifier or predicate; for example, *the parked car*. In this function the past participle has a passive meaning with transitive verbs, and a perfective meaning with unaccusative intransitive verbs; see Levin (1993, 86–88) for details. For reasons that are not clear to us, a predominant number of the high-frequency verbs cannot felicitously be used as prenominal adjectives. These verbs include unergative intransitives like *walk*, for which one would not expect to find the adjectival usage, given the above characterization; but they also include clear transitives like *move*, *try*, and *ask*, and unaccusative intransitives like *appear*, which are not generally felicitous in this usage. Consider: *?a moved car*, *?a tried approach*, *?an asked question*, *?an appeared ad*; but contrast: *an oft-tried approach*, *a frequently asked question*, *a recently appeared ad*, where an adverbial modifier renders the examples felicitous.[5] Among the low-frequency verbs, including *accentuate*, *bottle* and *incense*, the predominate types are those in which the past participle usage is preferred. What is clear from the plot in the top panel of Figure 2 is that the downward trend in the regression curve to the right of the plot is due to the lexical properties of a relatively small number of high-frequency verbs. For the greater part of the frequency range, there is a relatively stable proportion of participles to finite past forms. Thus, the hapax-based MLE yields an estimate that is uncontaminated by the lexical properties of individual high-frequency forms.

Finally, consider the Dutch verb forms *-en* that we started with. In Figure 1 the strong downward trend in the regression curve at the right of the figure is due in large measure to the inclusion of high-frequency auxiliary verbs, examples of which have already been given. These verbs, while possible in the infinitival form, occur predominantly in the finite form. Hence, a form such as *hebben* 'have' is much more likely to be a plural finite form than it is to be an infinitive. At the low end of the frequency spectrum, we find a great many verbs derived with separable particles, such as *afzeggen* 'cancel'; note that separable prefixation is the most productive verb-forming process in Dutch. In the infinitival form, the particle is always attached to

[5] One reviewer has suggested that the infelicity of many adjectival passives relates to the fact that the action denoted by the base verb is not regarded as producing an enduring result that affects the object denoted by the (deep) internal argument: contrast *a broken vase*, where the vase is enduringly affected by the breaking, with *?a seen movie*, where the movie is not affected. However, this cannot be the whole story since the object denoted by the internal argument of *kill* is presumably enduringly affected by the killing, yet *?a killed man* seems about as odd as *?a seen movie*.

the verb. However, in the finite forms in main clauses, the particle must be separated: for example, *wij* **zeggen** *onze afspraak* **af** 'we are cancelling our appointment'. These properties of Dutch separable verbs boost the likelihood of infinitival forms for the low-frequency ranges, but they also boost the likelihood of (higher-frequency) finite plural forms such as *zeggen*: since the separated finite plural form *zeggen* is identical to the finite plural of the underived verb *zeggen* 'say', any separated finite forms will accrue to the frequency of the generally much more common derivational base.

What all of these cases share is that the statistical properties of the high-frequency ranges are dominated by lexical properties of particular sets of high-frequency words. This in turn biases the overall MLE and makes it a poor predictor of novel cases. For example, auxiliaries such as *hebben* 'have' are among the most common verbs in Dutch, but they have rather different syntactic, and hence morphological, properties from other verbs; these properties in turn contaminate the high-frequency ranges and thus the overall MLE. In contrast, words in the low-frequency ranges, and particularly hapaxes, are heavily populated with (necessarily non-idiosyncratic) neologisms derived via productive morphological processes (Baayen 1989; Baayen and Renouf 1996). Any lexical biases that are inherent in these morphological processes — for example, the fact that a low frequency Dutch word ending in *-en* is more likely to be a noun than a verb — are well-estimated by the hapaxes. Now, for a sufficiently large training corpus, we can be very confident that an *unseen* complex word is non-idiosyncratic and formed via a productive morphological process, and this confidence increases as the corpus size increases (Baayen and Renouf 1996). Since the hapaxes of a particular morphological process mostly consist of non-idiosyncratic formations from that process, it makes sense that the distribution of a property among the hapaxes is the least contaminated estimate available for the distribution of that property among the unseen cases.

The hapax-based MLE that we have proposed is not only observationally preferable to the overall MLE, it is also firmly grounded in probability theory. The probability of encountering an unseen word given that this word is a word in *-en* is estimated by:

$$\Pr(\text{unseen}|\textit{-en}) \approx \frac{N_{1,N}(\textit{-en})}{N(\textit{-en})}, \qquad (1)$$

where $N_{1,N}(\textit{-en})$ denotes the number of hapax legomena in *-en* among the $N(\textit{-en})$ tokens in *-en* in the training sample; see Baayen (1989), Baayen and Lieber (1991), Good (1953), and Church and Gale (1991). Of course, this estimate is heavily influenced by the highest-frequency words in *-en*, as these words contribute many tokens to $N(\textit{-en})$. In our example, high-frequency auxiliaries such as *hebben* cause the probability of sampling unseen types in *-en* to be low — newly sampled tokens have a high probability of being an auxiliary rather than some previously unseen word. Interestingly, (1) can be used to derive an expression for the conditional probability that a word is, say, a noun, given that it is an unseen type in *-en* (Baayen 1993):

$$\begin{aligned}\Pr(\text{noun} \mid \text{unseen } \textit{-en} \text{ type}) &= \frac{\Pr(\text{noun} \cap \text{unseen } \textit{-en} \text{ type})}{\Pr(\text{unseen } \textit{-en} \text{ type})} \qquad (2)\\ &\approx \frac{\frac{N_{1,N}(\textit{-en}, \text{noun})}{N(\textit{-en})}}{\frac{N_{1,N}(\textit{-en})}{N(\textit{-en})}}\\ &= \frac{N_{1,N}(\textit{-en}, \text{noun})}{N_{1,N}(\textit{-en})}.\end{aligned}$$

Note that the estimator exemplified in (1) has been applied twice: once (in the denominator) to the distribution of all *-en* words; and once (in the numerator) to the distribution of the *-en* nouns — after reclassifying all verbal tokens in *-en* as representing one (very high-frequency) noun type in the frequency distribution. Similarly, the probability that an unseen word in *-en* is a verb is given by

$$\Pr(\text{verb} \mid \text{unseen } \textit{-en} \text{ type}) \approx \frac{N_{1,N}(\textit{-en}, \text{verb})}{N_{1,N}(\textit{-en})}. \tag{3}$$

Thus the proportion of verbal hapaxes in *-en* that we have suggested as an adjusted MLE estimator on the basis of the curve shown in Figure 2 is in fact an estimate of the conditional probability that a word is a verb, given that it is an unseen type in *-en*.

The results of the analyses presented in this paper are of potential importance in various applications that require lexical disambiguation and where an estimate of lexical priors is required. For high-frequency words, one can obtain fairly reliable estimates of the lexical priors by tagging a corpus that gives a good coverage to words of various ranges. For predicting the lexical priors for the much larger mass of very low-frequency types, most of which would not occur in any such corpus, the results we have presented suggest that one should concentrate on tagging a good representative sample of the hapaxes, rather than extensively tagging words of all frequency ranges.

Acknowledgments

The authors wish to thank four anonymous reviewers for *Computational Linguistics* for useful comments on this paper.

References

Baayen, Harald. 1989. *A Corpus-Based Approach to Morphological Productivity: Statistical Analysis and Psycholinguistic Interpretation.* Ph.D. thesis, Free University, Amsterdam.

Baayen, Harald. 1993. On frequency, transparency and productivity. *Yearbook of Morphology 1992,* pages 181–208.

Baayen, Harald and Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics,* 29:801–843.

Baayen, Harald and Antoinette Renouf. 1996. Chronicling the Times: Productive lexical innovations in an English newspaper. *Language,* 72:69–96.

Beard, Robert. 1995. *Lexeme-Morpheme Base Morphology.* SUNY, Albany.

Church, Kenneth. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing,* pages 136–143, Morristown, NJ. Association for Computational Linguistics.

Church, Kenneth Ward and William Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language,* 5(1):19–54.

Cleveland, William. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the Acoustical Society of America,* 74(368):829–836, December.

DeRose, Stephen. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics,* 14:31–39.

Francis, W. Nelson and Henry Kucera. 1982. *Frequency Analysis of English Usage.* Houghton Mifflin, Boston.

Good, I. 1953. The population frequencies of species and the estimation of population parameters. *Biometrica V,* 40(3,4):237–264.

Kupiec, Julian. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language,* 6:225–242.

Levin, Beth. 1993. *English Verb Classes and Alternations.* University of Chicago, Chicago.

Uit den Boogaart, P. C., editor. 1975. *Woordfrequenties in Gesproken en Geschreven Nederlands.* Oosthoek, Scheltema and Holkema, Utrecht.

Yarowsky, David. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of COLING-92,* Nantes, France, July. COLING.

Yarowsky, David. 1994. Decision lists for lexical ambiguity resolution. In *Proceedings of the 32nd Annual Meeting.* Association for Computational Linguistics.