7

Affix Discovery based on Entropy and Economy Measurements

ALFONSO MEDINA-URREA

This paper briefly describes an entropy and economy-based word segmentation method. Results of its application to discover items belonging to affix subsystems of two unrelated American languages are presented; namely, a variant of Ralámuli or Tarahumara (Uto-Aztecan) and one of Chuj (Mayan). More importantly, an attempt is made to compare these experiments in order to evaluate this approach by means of precision and recall measurements.

The following sections (7.1, 7.1.1, 7.1.2) present the method. Data obtained from the experiments is shown and discussed in section 7.2. The evaluation is presented in section 7.3.

7.1 Method

There are several prominent approaches to word segmentation. The earliest one is due to Zellig Harris, who first examined corpus evidence for the automatic discovery of morpheme boundaries for various languages, Harris (1955). His approach was based on counting phonemes preceding and following a possible morphological boundary: the more variety of phonemes, the more likely a true morphological border occurs within a word. Later, Nikolaj Andreev designed in the sixties the first automatic method based on character string frequencies which applied to various languages. His work was oriented towards the discovery of whole inflectional paradigms and applied to Russian and several other languages, Cromm (1996); and that of Kock and Bossaert (1974, 1978) in the seventies for French and Spanish. More recent prominent approaches deal with bigram statistics, see for instance Kageura (1999);

Texas Linguistics Society 10.
Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer and Elias Ponvert eds.

minimal distance methods, Goldsmith (2001); and morphotactics, Creutz and Lagus (2005).

The approach proposed in this paper grades word substrings according to their likelihood of representing a true affix or valid sequence of affixes. The resulting candidates are gathered in a table for later evaluation by experts. In essence, two quantitative measurements are obtained for every possible segmentation of every word found in a corpus: Shannon's entropy and a measure of sign economy (which will be dealt with, respectively, in sections 7.1.1 and 7.1.2). In short, the highest averaged values of these two measurements are good criteria to include word fragments as items in the table which will be called affix *catalog*, *i.e.* a list of affix candidates and their entropy and economy normalized measurements, ordered from most to least affixal.

The idea behind using these two measurements has to do with the preliminary notion that their combination may give a good quantitative estimate of how morphemes adhere to each other across languages. If Edward Sapir conceived of an energy among morphemes which glues them to each other throughout time to constitute linguistic structures, Sapir (1921), these two measurements can perhaps provide an estimation of this glutinous force or glutinosity in terms of bits carried by economical structure.

Regarding other approaches, this one differs from minimal distance methods which seek to find *the best* morphological model, Goldsmith (2001). An important goal of segmenting words is the discovering of affix paradigms. However, there are many kinds of paradigms. In fact, the better organized ones coexist with the lesser organized paradigms. And, even though these may be traced by automatic means, minimal distance methods prefer, by definition, the more compact paradigm types. In order to avoid exclusion of valuable items, the method sketched below gathers every possible candidate in a format that can later be used for human evaluation.

Another important approach focuses on the morphotactics of discovered items and is due to Creutz and Lagus (2005). This approach relies on morph probabilities and some meaning features (mainly, intraword right and left perplexity), which are roughly parallel to economy and entropy measurements. Although their work is more complete in that it proposes the morphotactics for each word, the present paper relies on a notion of linguistic sign economy, which seems to me much more appropriate than probabilities, see Medina-Urrea (2000). Needless to say, both approaches should be further examined and tested.

In fact, interesting comparison of methods exists for various languages, see Hafer and Weiss (1974), Kageura (1999), Medina-Urrea (2000), among others. But new comparison experiments must be conducted, that take into account more languages and the very diverse objectives that morphological segmentation may have, including those requiring least productive morphemes not to be excluded.

7.1.1 Information Content

High entropy measurements have been reported repeatedly as more or less successful indicators of borders between bases and affixes, see Hafer and Weiss (1974), Frakes (1992), Oakes (1998), Medina-Urrea (2000), Medina-Urrea and Buenrostro-Díaz (2003), Medina-Urrea and Hlaváčová (2005). These measurements are relevant because, as it was pointed out as early as the fifties by linguists like Joseph Greenberg (1967), shifts of amounts of information can be expected to correspond to the amounts of information that a reader or hearer would be bound to obtain from a text or spoken discourse. Frequent segments contain less information than those occurring rarely. Hence, affixes attach to those segments of a text (or discourse) which contain the highest amounts of information.

Information content of a set of word fragments is typically measured by applying Shannon's method.[1] Thus, for these experiments, the task was to measure the entropy of all word fragments which follow a prefix candidate and of those preceding each suffix candidate: borders between affixes and stems exhibit peaks of entropy. Specifically, looking for peaks of information meant taking each left-hand substring of each graphical word of the corpus, determining the probability of everything that follows, and applying Shannon's formula to obtain an entropy measurement for the right-hand substrings. Similarly, peaks of information were searched taking each right-hand substring of each word, determining the probability of everything that precedes it, and measuring entropy of the left hand substrings related to each right-hand fragment examined. Needless to say, this can be accomplished by means of two simple tree structures: one to store graphical words from left to right, the other to store them from right to left.

7.1.2 Economy Principle

Essentially, an economy measurement should represent how much linguistic structure there is in a given expression. If natural languages are systems, they and their components must be economical to some degree. Thus, we can expect certain signs to be more economical than others because they relate to other signs in an economical way. One aspect of sign economy is evident in that a sign at one level of language, say the lexical one, may be composed of more than one sign of the lower level, say the morphological one. In this

[1] Recall the formula

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{7.1}$$

where p_i stands for the relative frequency of word fragment i; Shannon and Weaver (1949); for brief descriptions see, among many others, Oakes (1998) or Manning and Schütze (1999).

manner, a language can refer at the lexical level to a great number of things using considerably fewer signs than it would be necessary if it had exactly one sign for each thing named, see Kock and Bossaert (1974, 1978).

From the syntagmatic perspective, affixes can combine with bases to produce a virtually infinite number of lexical signs. It is clear that affixes do not combine with every base. Certain ones combine with many bases, others with only a few. Nevertheless, it makes sense to expect more economy where more combinatory possibilities exist.

Regarding the paradigmatic dimension, affixes substitute other affixes, *i.e.* appear in complementary distribution in a corpus, when attaching to specific bases. If there is a relatively small set of alternating signs (paradigms) which adhere to a large set of unfrequent signs (to form syntagms), the relations between the former and the latter must be considered even more economical. This is naturally pertinent for both derivation and inflection.

The economy of segmentations can be measured for each word fragment (affix candidate) by comparing the sizes of two sets of word substrings. These sets can be described as follows:

1. *companions* — word fragments which appear attached to the affix candidate (syntagmatic relation).

2. *alternants* — word fragments which occur in complementary distribution with the affix candidate.

The following fraction is a simplified example of how these can be compared to capture the essence of the method, first proposed by Kock and Bossaert (1974, 1978) and later paraphrased by Medina-Urrea (2000, 2003):

$$k = \frac{companions}{alternants} \tag{7.2}$$

More formally, let $a_{i,j} :: B_{i,j}$ represent a set of graphical words, where $B_{i,j}$ is the set of word endings which follow, according to a corpus, a left-hand word segment $a_{i,j}$ (which consists of the first jth letters of the ith word). Every member of $B_{i,j}$ occurs in complementary distribution with each other. Furthermore, let $B^{s}_{i,j}$ be the subset of $B_{i,j}$ consisting of the right-hand word fragments which are suffixes of the language in question. Alternatively, let $A_{i,j} :: b_{i,j}$ be a set of words, where $A_{i,j}$ is the set of word beginnings which precede the right-hand word fragment $b_{i,j}$. Let $A^{p}_{i,j}$, a subset of $A_{i,j}$, be the set of word beginnings which are prefixes of the language and occur in complementary distribution with the word fragment $a_{i,j}$. Two ways to estimate the economy of a segmentation between a set word beginnings and a set of word endings are:

$$k^p_{i,j} = \frac{|B_{i,j}| - |B^s_{i,j}|}{|A^p_{i,j}|} \tag{7.3}$$

$$k^s_{i,j} = \frac{|A_{i,j}| - |B^p_{i,j}|}{|B^s_{i,j}|} \tag{7.4}$$

Essentially, the numerators of 7.3 and 7.4 can be described as the sets of *companions* of, respectively, the word segments $a_{i,j}$ or $b_{i,j}$ and the denominators the set of word fragments in paradigmatic relation to, respectively, $a_{i,j}$ or $b_{i,j}$ (*alternants*), when these both are assumed affix candidates.

In this way, when an initial word fragment is given, a very large number of companions and a relatively small number of alternants yield a high economy value. Meanwhile, a small number of companions and a large one of alternants indicate a low economy measurement. In the latter case, the word fragment in question is not very likely to represent exactly a morpheme nor a sequence of them.

7.1.3 Entropy and Economy combined

Both entropy and economy, as described above, complement each other in order to estimate what I called above glutinosity, which, when dealing with affixes and bases, refers specifically to the affixality of word fragments. In fact, the values obtained for a given word fragment can be averaged or multiplied. For the experiments described below, they were normalized and averaged. That is, *affixality* was estimated here by means of the arithmetic average of the relative values of entropy and economy: $(\frac{h_i}{\max h} + \frac{k_i}{\max k}) * \frac{1}{2}$, where h_i stands for the entropy value associated to prefix candidate i; k_i represents the economy measurement associated to the same candidate; and $\max h$ returns the maximum quantity of h calculated for all prefixes (same idea for $\max k$).

The important fact is that the highest values (those expected to occur at the borders between prefixes and bases, and between bases and suffixes) are good criteria to include word fragments as items in the Catalog of Affixes.

7.2 Experiments

Affix catalogs of this sort were generated to look into the Ralámuli derivational suffixes and the Chuj verbal inflection subsystem. In this section, these are examined to determine how many candidates are true affixes or sequences of them.

7.2.1 Ralámuli Derivational Suffixes

Ralámuli or Rarámuri, better known as Tarahumara, is a Uto-Aztecan agglutinative language spoken in northern Mexico. Word formation is accomplished by means of suffixation. Stems are followed by derivational suffixes,

TABLE 1
Catalog of Ralámuli Suffixes (top 30 candidates)

RANK	SUFFIX	FREC.	ECONOMY	ENTROPY	AFFIXALITY
1.	∼ma	35	1.00000	0.88030	0.98050
2.	∼re	77	0.81100	0.86060	0.82370
3.	∼sa	33	0.93060	0.75590	0.77430
4.	∼ra	62	0.64610	0.85080	0.71940
5.	∼si	28	0.52570	0.83450	0.70340
6.	∼na	25	0.72240	0.79840	0.64410
7.	∼go	4	0.90650	0.64930	0.59000
8.	∼é	49	0.43580	1.00000	0.53400
9.	∼ame	51	0.30640	0.85910	0.47250
10.	∼gá	18	0.37810	0.61360	0.46550
11.	∼ka	19	0.28060	0.84130	0.45920
12.	∼á	67	0.31330	0.91950	0.45710
13.	∼ré	11	0.41020	0.73430	0.43780
14.	∼ga	50	0.28340	0.80650	0.42430
15.	∼a	281	0.18960	0.97250	0.42250
16.	∼ba	8	0.30220	0.74000	0.41880
17.	∼ayá	8	0.44320	0.57570	0.41110
18.	∼í	42	0.26480	0.80540	0.39070
19.	∼či	39	0.27510	0.74000	0.37260
20.	∼e	164	0.29100	0.64290	0.36210
21.	∼mi	4	0.30220	0.69910	0.35760
22.	∼áame	12	0.00000	0.90570	0.35350
23.	∼yá	20	0.11080	0.57420	0.34260
24.	∼i	139	0.10320	0.84220	0.32820
25.	∼ira	11	0.10990	0.79570	0.32350
26.	∼o	41	0.00000	0.96810	0.32270
27.	∼ne	3	0.40290	0.41950	0.32170
28.	∼wa	9	0.13430	0.74000	0.30730
29.	∼agá	6	0.20140	0.53640	0.30150
30.	∼sí	4	0.00000	0.53740	0.29820

and these by inflectional ones. Since Ralámuli has very little inflection, we applied the method to examine suffixes of a derivational nature. The experiment conducted is described in Medina-Urrea and Alvarado-García (2004).

The text sample represents the dialectal variant from San Luis Majimachi, Bocoyna, Chihuahua. For today's standards, this sample is extremely small, consisting of 3,584 word-tokens and 934 word-types. Table 1 exhibits Ralámuli's top 30 suffix candidates. These candidates are presented in the second column. The third column exhibits the number of word-types where the candidate came out as the best possible suffix of that word-type. The fourth and fifth columns contain the normalized measurements of economy and entropy. The last column exhibits the affixality index (the arithmetic average of the entropy and economy values). Finally, the first column shows the rank of the candidates according to this index: the lower the rank, the greater the affixality index.

Even though Ralámuli has few inflectional forms, the larger catalog exhibits more items containing inflectional material than were expected. This is because input texts are constituted by linguistic acts in the pragmatic act of narrating, so words appear inflected. Nevertheless, if inflectional suffixes are considered more affixal than derivational ones, it should not be surprising to find the four most prominent Tarahumara inflection affixes appearing at the top of the table: ∼*ma*, ∼*re*, ∼*sa*, and ∼*si*, which mark tense, aspect and mode.

Regarding nominal and verbal derivational suffixes, the 35 most prominent ones were identified previously. 25 of these occurred within the first 100 catalog entries (a recall measure of 71% within this limit). The other entries are chains of suffixes (including sequences of derivational and inflectional items) and residual forms. The examination of residual items was especially difficult. Questions about lexicalized affixes (possibly fossilized items) and about the relationship between syllable structure and affix status emerged. These matters remain to be revised by Ralámuli experts. Meanwhile, for evaluation purposes (see evaluation section below), entries with unexpected syllabic structure were not counted as acceptable suffixes nor valid chains of them.

The 10 derivational suffixes which did not appear anywhere in the catalog are essentially verbal derivational forms, or modifiers of transitivity or of some semantic characteristic of verbal forms. This might mean that the small sample used is more representative of nominal structures, rather than of verbal ones. It is worth stressing that a significant part of the known Ralámuli derivational system, essentially the nominal subsystem, was retrieved from a small unrepresentative set of texts.

7.2.2 Chuj Verbal Inflection System

Chuj belongs to the Mayan family of languages and it is spoken on both sides of the border between Mexico and Guatemala. The experiment conducted is described in Medina-Urrea and Buenrostro-Díaz (2003). This language is particularly interesting for the present paper because its verbal inflection system is constituted by both prefixes and suffixes. The prefix and suffix catalogs obtained measuring entropy and economy are shown respectively in Tables 3 and 5.

The text sample used is also very small (15,485 word-tokens, about 2,300 types). Given its reduced size and the fact that it is composed of only five narrations, it cannot properly be considered a balanced and representative corpus of the language.[2]

[2] Results are nevertheless interesting because, given her grammatical interests, Buenrostro put

TABLE 2
Paradigm of Chuj Verbal Inflection Prefixes

TENSE

RANK	SUFFIX	AFFIXALITY
7.	tz∼	0.74
2.	ix∼	0.86
24.	x∼	0.51
12.	ol∼	0.65
—	ø∼	—

GRAMMATICAL PERSON
ABSOLUTIVE

PERSON	RANK	SUFFIX	AFFIXALITY
1	1.	in∼	0.91
2	27.	ač∼	0.50
3	—	ø∼	—
1	74.	onh∼	0.33
2	251.	ex∼	0.20
3	—	ø∼ eb’	—

GRAMMATICAL PERSON
ERGATIVE

PERSON	RANK	SUFFIX	AFFXY.	RANK	SUFFIX	AFFXY.
1	1.	in∼	0.91	13.	w∼	0.63
2	8.	a∼	0.71	—	ø∼	—
3	3.	s∼	0.83	17.	y∼	0.56
1	5.	ko∼	0.77	58.	k∼	0.36
	43.	ku∼	0.42			
2	22.	e∼	0.54	183.	ey∼	0.24
3	3.	s∼ eb’	0.83	17.	y∼ eb’	0.56

Chuj Prefixes

Buenrostro’s proposal of Chuj’s verbal inflection prefixes appears in Table 2. Every item is listed with its rank to the left and with its affixality index to the right (both ranks and affixality indexes were obtained from the catalog partially shown in Table 3). Tense markers are shown in the first section. In the second and third sections, markers appear which indicate absolutive and ergative grammatical person. In the third section, ergatives to the right attach to vowel initial stems and those to the left attach to consonant initial ones.[3]

Regarding the prefix catalog (Table 3), all tense markers, most ergatives

special emphasis in compiling a collection of texts representative of verbal structures.

[3] Thus, *ko∼* and *ku∼* are allomorphs representing ergative, 1st person plural, which attach to consonant initial stems.

TABLE 3
Catalog of Chuj Prefixes (top 30 candidates)

RANK	PREFIX	FREC.	ECONOMY	ENTROPY	AFFIXALITY
1.	in∼	93	0.83990	0.98280	0.91130
2.	ix∼	181	0.80210	0.90880	0.85540
3.	s∼	187	0.66620	0.98740	0.82680
4.	kak'∼	1	1.00000	0.59290	0.79650
5.	ko∼	71	0.66030	0.87830	0.76930
6.	xsči'∼	1	1.00000	0.51070	0.75540
7.	tz∼	349	0.59450	0.88610	0.74030
8.	a∼	164	0.41110	1.00000	0.70550
9.	tzin∼	48	0.44820	0.93380	0.69100
10.	olin∼	26	0.47320	0.88180	0.67750
11.	xal∼	2	0.67500	0.66010	0.66750
12.	ol∼	185	0.52790	0.78070	0.65430
13.	w∼	70	0.73820	0.52560	0.63190
14.	olač∼	26	0.47210	0.76630	0.61920
15.	tzs∼	49	0.42520	0.81010	0.61770
16.	ixin∼	29	0.36740	0.81830	0.59290
17.	y∼	127	0.54740	0.57100	0.55920
18.	k'a∼	11	0.36360	0.74520	0.55440
19.	ma∼	31	0.22020	0.87220	0.54620
20.	al∼	15	0.21230	0.87400	0.54320
21.	na∼	9	0.30950	0.77610	0.54280
22.	e∼	63	0.16630	0.91730	0.54180
23.	ak'∼	12	0.23990	0.78810	0.51400
24.	x∼	43	0.25860	0.75680	0.50770
25.	tzonh∼	16	0.22490	0.78660	0.50570
26.	b'ati∼	1	0.75000	0.25540	0.50270
27.	ač∼	9	0.19440	0.80340	0.49890
28.	ixs∼	24	0.16880	0.81540	0.49210
29.	ay∼	23	0.43870	0.53450	0.48660
30.	k'e∼	3	0.25000	0.70220	0.47610

and a couple of the absolutives appear.[4] Finally, 10 residual entries also appear (verb stems, non-readily recognizable or fragmented prefixes).

Hence, precision for this table would be the proportion of right guesses, 66.6%. With respect to recall, every prefix listed among Buenrostro's set appeared in the prefix catalog among the first 251 entries of Table 3, which only displays the first 30. So there is a recall of 100% for the first 251 catalog items.

[4] Notice that there are chains of tense and person marker prefixes: *tz.in∼*, *ol.in∼*, *ix.in∼*, *ix.s∼*, *tz.s∼*, *tz.onh∼*, and *ol.ač∼*. In fact, personal markers are word initial because one of the tense markers is the null affix ø-.

TABLE 4
Paradigm of Chuj Verbal Inflection Suffixes

VOICE			
	RANK	SUFFIX	AFFIXALITY
passive	63.	∼aj	0.4129
	68.	∼chaj	0.4018
	872.	∼b'il	0.1212
	1016.	∼nax	0.0949
	—	∼**ji**	—
antipassive	19.	∼an	0.5958
	28.	∼wi	0.5531
	161.	∼waj	0.2629

MODAL/ TEMPORAL			THEMATIC VOWEL		
RANK	SUFFIX	AFFIXALITY	RANK	SUFFIX	AFFIXALITY
6.	∼ok	0.7479	11.	∼i	0.6703
18.	∼nak	0.5977	12.	∼a	0.6549

Chuj Suffixes

Table 4 shows Buenrostro's proposal for the Chuj's inflectional suffix subsystem. These suffixes mark voice, mode and end of utterance. Thematic vowels distinguish transitive from intransitive verbs and signal end of phrase. Again, items appear surrounded by their ranks and affixality values, copied from the suffix catalog partially shown in Table 5. Almost all items proposed by Buenrostro occur within the first 1016 entries of the suffix catalog.[5] This corresponds to a recall of 92% (eleven of twelve) and of 75% for the first 500 entries (nine of twelve). Taking prefixes and suffixes together, recall would be of 96.55% (28 items of 29) for the first 1016 items, and 91% within the first 500 catalog entries (29 of 32).

7.3 Evaluation

For the sake of simplicity and since the experiments described above deal with specific affix subsystems, such as inflectional or derivational, prefixal or suffixal, I will base the following considerations on their size. Specifically, a window of the size of each relevant subsystem to look into the top of the relevant affix catalog was used to calculate precision and determine the proportion of errors. Notice that items belonging to other subsystems also appear within that window. They were not counted as errors.

Furthermore, a recall measure deals specifically with how much of the sub-

[5] Except passive voice marker ∼*ji*, which is shown in boldface.

TABLE 5
Catalog of Chuj Suffixes (top 30 candidates)

RANK	SUFFIX	FREC.	ECONOMY	ENTROPY	AFFIXALITY
1.	∼kan	68	1.00000	0.90290	0.95150
2.	∼nej	24	0.98110	0.76980	0.87540
3.	∼ta’	70	0.75260	0.82030	0.78650
4.	∼b’at	63	0.67170	0.86590	0.76880
5.	∼al	82	0.53560	1.00000	0.76780
6.	∼ok	68	0.55850	0.93740	0.74790
7.	∼ab’	49	0.50780	0.90590	0.70690
8.	∼il	62	0.46340	0.93470	0.69900
9.	∼ač	16	0.71490	0.67950	0.69720
10.	∼xi	37	0.70030	0.68290	0.69160
11.	∼i	205	0.46370	0.87690	0.67030
12.	∼a	142	0.37910	0.93060	0.65490
13.	∼kot	48	0.55640	0.74040	0.64840
14.	∼el	68	0.43240	0.86430	0.64830
15.	∼tak	19	0.37850	0.90620	0.64240
16.	∼in	46	0.34980	0.89170	0.62070
17.	∼kani	8	0.48810	0.71660	0.60240
18.	∼nak	18	0.41430	0.78120	0.59770
19.	∼an	233	0.36460	0.82710	0.59580
20.	∼alan	13	0.36500	0.82250	0.59380
21.	∼ab’i	9	0.54200	0.64330	0.59260
22.	∼ni’	7	0.61660	0.56680	0.59170
23.	∼k’oč	28	0.29220	0.86700	0.57960
24.	∼ak’	43	0.34370	0.79220	0.56800
25.	∼ak’tej	6	0.52520	0.58790	0.55660
26.	∼koti	18	0.42930	0.68360	0.55640
27.	∼ila	9	0.54070	0.57080	0.55580
28.	∼wi	14	0.42080	0.68530	0.55310
29.	∼ik’	12	0.48130	0.60150	0.54140
30.	∼o	123	0.11440	0.96340	0.53890

system sought is not retrieved. In this case, a larger evaluation window was used, because these languages have other subsystems competing to appear towards the beginning of the catalog, so items of one subsystem appear mixed with those of other complex subsystems. Upon examination of results, a window of 500 was selected (the Ralámuli catalog has fewer items than that). Obviously, a smaller window means greater precision and lower recall, whereas a greater window means lower precision and greater recall. Therefore, different window sizes for precision and recall will maximize both measurements —the smaller window (subsystem size) for precision, the bigger one (of 500) for recall. It should be clear that precision will decrease considerably as the window grows because rarer items are mixed with plain mistakes and unrecognizable, residual forms (which were counted as errors). Conversely, recall

TABLE 6
Evaluation measurements

	Ralámuli	Chuj	
subsystem	derivational suffixes	prefix verbal inflection	suffix verbal inflection
sample tokens	3,584	15,485	
sample types	934	2,300	
subsystem size[a] (n)	35	20	12
right guesses[b]	28	15	12
presumed errors[b]	7	5	0
unretrieved items[c]	10	0	3
precision	0.80	0.75	1.00
recall	0.71	1.00	0.75

[a] Allomorphs, homographs and polysemous items count separately; null affixes are excluded.
[b] In relevant catalog within subsystem size.
[c] Members of subsystem not found in relevant catalog within first 500 catalog candidates.

will decrease as the window is diminished because rare and lesser productive members of the subsystem examined will fall outside the smaller window (whereas other subsystem's more productive members also compete for the upper catalog entries).

Table 6 shows numerical data, as well as precision and recall measurements. The first line shows the name of the affix subsystem focused. The second and third lines characterize the corpora used: their size in number of word-tokens and number of word-types. The fourth line shows the size of the subsystem sought in those corpora.

It is worth stressing that determining subsystem size is indeed a problem not to be underestimated. For each of the experiments reported here, specialists studied the subsystem in order to know its size, which surely varies from perspective to perspective. Here, null morphemes were excluded and allomorphs, homographs, and polysemous items were counted separately.

Then, based on a window of the size of the subsystem in focus for each corpus, the correct guesses and presumed errors were counted within that window (lines fifth and sixth). The seventh line contains the number of subsystem members not found within the much larger window of five hundred catalog items. The last two lines exhibit the precision and recall measures. As mentioned above, precision is the proportion of correct guesses within the first n entries of the relevant catalog, such that n is the size of the subsystem sought. Additionally, recall is the number of members of the subsystem actually found within the first 500 hundred catalog items.

The evaluation measures look good, in part because window sizes were selected to maximize them. Nevertheless, there is a sense in which affixal items tend to be concentrated towards the top of their catalogs and that most of the subsystems in which the items play a part can be retrieved within the first five hundred catalog entries.

7.4 Final Remarks

In this paper, results of applying a method for affix discovery to two American languages represented by very small corpora were presented and evaluated. In spite of all possible improvements, the method has already yielded sets of items that can be used for the development of other tools for the computational processing of these languages, such as stemmers, lemmatizers and morphological analizers. It seems relevant to point out that the plain lists of affix candidates examined cannot be considered *the* morphological models of the focused languages. They are rather windows to complex phenomena which can be described in different ways, according to the preferred language theoretical perspective. They are tools for the discovery of the unknown, more related to text mining than to rule-based formalism design.

However, the values obtained for the morphological items estimate the extent to which these glue to their bases. They represent a measurement of how much they agglutinate to form words and to inflect them. It will be very interesting to measure this glutinous force or glutinosity at all points of the word where morpheme boundaries occur in order to improve any possible study of the morphotactics of the languages examined. Also, it would be worthwhile to find out whether this Sapirean energy can be best estimated by means of some other scheme. Meanwhile, the one presented, which measures bits carried by economical signs, seems to be simple, and appropriate enough.[6]

Acknowledgements

I am very thankful to Cristina Buenrostro and Maribel Alvarado for their invaluable collaboration. This work was financed by a UNAM's grant DGAPA PAPIIT IN400905. Also, I thank the support of the USI-IINGEN and the UNAM's *Macroproyecto de Tecnologías para la Universidad de la Información y la Computación del Programa Transdisciplinario en Investigación y Desarrollo*.

References

Creutz, M. and K. Lagus. 2005. Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*. Espoo, Finland.

[6] A set of axioms defining such measurement scheme is proposed in Medina-Urrea (2003).

Cromm, O. 1996. Affixerkennung in deutschen Wortformen. Eine Untersuchung zum nicht-lexikalischen Segmentierungsverfahren von N. D. Andreev. Abschluß des Ergänzungsstudiums Linguistische Datenverarbeitung, Frankfurt am Main.

Frakes, W. B. 1992. Stemming Algorithms. In W. B. Frakes and R. Baeza-Yates, eds., *Information Retrieval, Data Structures and Algorithms*, pages 131–160. New Jersey: Prentice Hall.

Goldsmith, J. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27(2):153–198.

Greenberg, J. H. 1967. *Essays in Linguistics*. Chicago: The University of Chicago Press.

Hafer, M. A. and S. F. Weiss. 1974. Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval* 10:371–385.

Harris, Z. S. 1955. From Phoneme to Morpheme. *Language* 31(2):190–222.

Kageura, K. 1999. Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences. *Journal of Quantitative Linguistics* 6:149–166.

Kock, J. de and W. Bossaert. 1974. *Introducción a la lingüística automática en las lenguas románicas*, vol. 202 of *Estudios y Ensayos*. Madrid: Gredos.

Kock, J. de and W. Bossaert. 1978. *The Morpheme. An Experiment in Quantitative and Computational Linguistics*. Amsterdam, Madrid: Van Gorcum.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Medina-Urrea, A. 2000. Automatic Discovery of Affixes by Means of a Corpus: A Catalog of Spanish Affixes. *Journal of Quantitative Linguistics* 7(2):97–114.

Medina-Urrea, A. 2003. *Investigación cuantitativa de afijos y clíticos del español de México. Glutinometría en el Corpus del Español Mexicano Contemporáneo*. Ph.D. thesis, El Colegio de México, Mexico.

Medina-Urrea, A. and M. Alvarado-García. 2004. Análisis cuantitativo y cualitativo de la derivación léxica en ralámuli. In *Primer Coloquio Leonardo Manrique*. Conaculta-INAH, Mexico.

Medina-Urrea, A. and E. C. Buenrostro-Díaz. 2003. Características cuantitativas de la flexión verbal del chuj. *Estudios de Lingüística Aplicada* 38:15–31.

Medina-Urrea, A. and J. Hlaváčová. 2005. Automatic Recognition of Czech Derivational Prefixes. In *Proceedings of CICLing 2005*, vol. 3406 of *Lecture Notes in Computer Science*, pages 189–197. Berlin/Heidelberg/New York: Springer.

Oakes, M. P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Sapir, Edward. 1921. *Language: An Introduction to the Study of Speech*. New York: Harcourt, Brace & Company.

Shannon, C. E. and W. Weaver. 1949. *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.