American Journal of Computational Linguistics *Microfiche 36 : 52*

A Natural Language Processing Package

David Brill and Beatrice T. Oshika

Speech Communications Research Laboratory, Inc.
800A Miramonte Drive
Santa Barbara, California 93109

ABSTRACT

A set of SAIL programs has been implemented for analyzing large bodies of natural language data in which associations exist between strings and sets of strings. These programs include facilities for compiling information such as frequency of occurrence of strings (e.g. word frequencies) or substrings (e.g. consonant cluster frequencies), and describing relationships among strings (e.g. various phonological realizations of a word). Also, an associative data base may be interactively accessed on the basis of keys corresponding to different types of data elements, and a pattern matcher allows retrieval of incompletely specified elements. Applications of this natural language processing package include analysis of phonological variation for specifying and testing phonological rules, and comparison across languages for historical reconstruction.

I. NATURAL LANGUAGE PROCESSING PACKAGE

A. General characteristics

The natural language processing package implemented at the Speech Communications Research Laboratory (SCRL) is currently used in the analysis of associated lists of string data such as discourse transcriptions or pronouncing dictionaries. The package consists of

a) a set of "batch" programs which provide frequency and context information on the lexical and phonological forms appearing in the input; and

b) a system for interactively accessing the data on the basis of orthographic and phonological patterns.

All of the programs in this package are written in SAIL, an ALGOL-based language offering extended string and set manipulation operations and an associative data base. The programs run on a DEC PDP-10 at Carnegie-Mellon University via the Advanced Research Projects Agency (ARPA) computer network (ARPANET). The ARPANET is accessed by the ELF operating system developed by SCRL, which runs on a local PDP-11 [1].

While the processing package is applicable to various types of natural language data, it has been used most extensively at SCRL in the analysis of discourse transcriptions. The discourses consist of conversational speech gathered in interviews with adult speakers of various dialects of American English. More than twenty-five discourses, transcribed orthographically and phonologically, have been processed, yielding

detailed information on over 28,000 utterances representing about 3,500 distinct lexical items. All examples in this section are taken from a typical discourse.

B. "Batch" Facility

Discourse processing usually begins with the generation of a transcription reference file in which orthographic and phonological representations are listed in discourse order, as illustrated in Figure 1.

897	WELL	885	//WELEHTS//
898	LET'S		
899	TRY	886	TRAY
900	CLASSIFYING	887	KLAES$CFAYIHNX
901	THEM	888	DHAXM
902	ACCORDING=	889	//AXK$ORDIHN/TUW//
903	TO		
904	THE	890	DH$I
905	EXCUSES	891	IHKSYUWS$CZ

Figure 1

In this example, the phonological realization of TRY is /tray/ (coded TRAY). The phonological code shown is a basic ARPA phonemic alphabet augmented by special symbols indicating some phonetic detail, such as vowel height. The realization of THE, for example, is coded DH$I, indicating that the vowel fell between /i/ and /ɪ/.

Reference numbers assigned to each utterance serve as an index to the discourse context in which utterances occur, and are used to interpret the output of other programs in the package. Separate reference number sequences are provided for

the orthographic and phonological forms in the reference files, since there may not be a one-to-one correspondence between these forms, as in the case of phonological merging which obscures word boundaries. In Figure 1, for example, the two orthographic items WELL and LET'S are realized as a single phonological item /wlɛts/ (coded WELEHTS)

The core of the "batch" processing facility is a set of three programs: PROCON, ENVIRN and CLUSTR. PROCON provides frequency and context information on the lexical level, while the other two provide similar information on the phonological level.

PROCON output contains an alphabetically sorted list of the utterance types occurring in the input discourse transcription file as illustrated in Figure 2. Frequency of occurrence of each type is given, along with the various phonological realizations. For each phonological realization, frequency count and reference numbers are provided.

8	HAVE	3	AXV	11,337,703
		3	HHAE\	354,828,1397
		1	HHAXV	710
		1	HH$GV	1067

Figure 2

In Figure 2, for example, HAVE occurred eight times, and was pronounced AXV (/əv/) three times and HHAEV (/hæv/) three times. Using the reference numbers associated with these pronunciations, it is possible to establish the discourse context.

One would find that the three AXV pronunciations (i.e. utterances 11, 337 and 703) all involved the auxiliary construction in "...may have felt...seemed to have been ..which have since been..."

ENVIRN tallies occurrences of phonological segments and environments in the discourse transcriptions. The output of this program lists frequencies of all phonemes appearing in the input file, as illustrated in Figure 3.

```
Q  30

D--EN       1     486
EH--EN      8     189, 200, 223, 226, 233,
                  248, 368
ER--EN      2     1427, 1444
EY/--Y      1     1361
IH--/DH     1     134
IH--/IH     1     1416
IH--/S      1     120
K/--/DH     1     1178
```

Figure 3

Glottal stop, coded Q, occurred a total of thirty times in the discourse. The immediate environments of Q are listed alphabetically by left context, with word boundaries indicated by slash /, and a frequency count and reference numbers are given for each environment. For example, Q appeared eight times in the context EH--EN (/ɛ--n/), and a check of the reference list shows that all these occurrences were in the word sentence(s).

ENVIRN output also provides a frequency ordered list of phonemes, with frequency totals broken down according to occurrence in word initial, medial and final position.

CLUSTR, the third of the "batch" programs,is used in the analysis of phoneme cluster distribution in the discourse data. All clusters are indexed by each of their component phonemes, so that the cluster NDZ (/ndz/) which is listed under D in Figure 4 also appears under N and Z in the full output.

D 70

B D	1	169
D EN	2	71, 103
D EN T	2	593, 1127
D EN T S	1	699
D Q EN T S	1	486
D V	2	1417, 1445
D Z	5	278, 284, 837, 1341, 1350
L D	12	38, 385, 616, 712, 923, 1248, 1465, 1474, 1478, 1480, 1494, 1512
M D	1	330
W D	32	35, 118, 186, 227, 328, 400, 419, 550, 608, 608, 627, 631, 653, 670, 682, 704, 717, 730, 730, 745, 853, 933, 1039, 1194, 1199, 1201, 1228, 1233, 1253, 1320, 1372, 1425
N D Z	1	1429
R D	5	630, 889, 1277, 1303, 1314
V D	3	187, 750, 765
Z D	2	106, 451

Figure 4

Separate output may be generated for clusters occurring within words or across word boundaries. Currently, consonant and vowel clusters are tallied, but the program can be easily modified to handle sequences of phonemes belonging to arbitrary user-defined classes (e.g. voiced sounds,, nasals, unvoiced stops, etc.).

For each phoneme belonging to a selected class, CLUSTR provides a count of the number of times that the phoneme appears in clusters, an alphabetically sorted list of those clusters, and a frequency count and reference numbers for each cluster. Figure 4, a sample of CLUSTR output for within-word consonant clusters, shows that D appeared in clusters a total of 70 times, with 32 of these being ND clusters. Reference numbers may be used to establish the discourse context of any cluster. For example, the cluster D Q EN T S (/dʔnts/) appears in utterance 486 which is the word students. Like ENVIRN, CLUSTR provides a frequency ordered list of cluster types in addition to the alphabetic list.

C. Interactive Retrieval Facility

The set of "batch" programs is complemented by a language data retrieval system which allows the user to interactively retrieve data items conforming to various orthographic, phonological and syntactic patterns.

Linguistic data is internally stored in the system as a network of associations between items of various types. These associations are implemented in SAIL as LEAP triples [2] and the element types entering into these associations vary accord-

ing to the particular application. For example, in analysis of the discourse data described above, triples contain orthographic, phonological and syntactic elements. For study of phonetic-to-phonemic mapping, triples might be orthographic, phonemic and phonetic elements. In comparative linguistic research, triples might consist of an orthographic element and two phonological elements corresponding to two languages or dialects

Data can be accessed on the basis of patterns directed to any one (or any combination) of these elements. For example, if the data base contains associations between orthographic, phonological and syntactic elements, then the query

P/ O: THE

retrieves the phonological items associated with the spelling THE, and might return DHAX (/ðə/) and DHIY (/ðiy/). The query

O/ P: TUW

would return the orthographic items pronounced TUW (/tu/), e.g <u>two</u>, <u>too</u>, <u>to</u>.

Patterns such as THE and TUW completely specify the element to which they are directed, but various special forms allow partial specifications to be expressed also. The symbol $ matches any single segment (in a phonological pattern) or character (in an orthographic pattern), and the symbol = matches any number, including zero, of contiguous segments (or characters). Thus, if N is the syntactic code for Noun, the query

O/ P: $$, S: N, O: D=

searches for all two-phoneme nouns which begin with the letter D, and might return dye, day, doe, dough.

Each phonological element is defined in terms of a set of features such as UV (unvoiced) and ST (stop), and these features may be used to specify segments in phonological patterns. To search for phonological realizations containing /i/ between unvoiced stops, one could use the query

P/ P: =⟨UV + ST⟩IY⟨UV + ST⟩=

to find /kip/ (keep), /pikɪŋ/ (peeking), and /rɪpit d/ (repeated)

Boolean operators are also available for specifying pattern segments. For example, the query

O/ O: (C OR K)=, P: (NOT K)=

returns orthographic items which begin with C or K and are not pronounced with initial /k/, e.g. cite, change, know.

Several capabilities lacking in the current interactive system will be available in the near future. The user will be able to (1) specify optional segments and sequences of segments in phonological patterns; (2) create and name sets containing items of interest, e.g. monosyllabic function words, and use set operations such as union and intersection; (3) interactively modify feature definitions of phonological symbols; (4) retrieve several elements, e.g. orthographic and phonological forms, simultaneously; (5) display the discourse context of any given item, and (6) write retrieval queries and responses to a file for subsequent analysis.

II. APPLICATIONS

The processing package can be used in the analysis of various kinds of natural language data, as illustrated in the following examples.

A. Phonological variation

The programs can be used to efficiently index and sort natural language data so that systematic phonological variation can be easily examined. For example, inspection of a PROCON output for a ten minute interview consisting of over 2,000 utterance tokens yields general observations such as

-- final /t/ alternates with final glottal stop /ʔ/ under certain conditions;

-- alveolar flapping occurs under several stress conditions which appear to be related to noun affixes.

These preliminary observations can be systematically investigated using the interactive query system.

The data base can be queried for all phonological realizations ending in T (/t/) or Q (/ʔ/), and the corresponding orthographic entries, using the queries

P/ P: =(T OR Q) and O/ P: =(T OR Q)

The resulting list might include

art	/ɑrt/	limit	/lɪmɪt/
but	/bət/		/lɪmɪʔ/
	/bəʔ/	raft	/ræft/
can't	/kænt/	that	/ðæt/
	/kᵊnʔ/		/ðɛʔ/
fished	/fɪʃt/	want	/wɑnt/
it	/ɪt/		/wɑnʔ/
	/ɪʔ/		

That is, final /t/ appears to vary with final /ʔ/ following vowels and following nasals, but not elsewhere. This hypothesis, represented as a context-sensitive phonological rule, could then be tested against additional data using any of several computer rule testers [3-5].

Forthcoming modifications will allow queries with set operations, such that the intersection of orthographic entries having final /t/ alternating with /ʔ/ can be requested directly by the query

O/ P: =T ∩ Q/ P: =Q .

That is, only entries with /t/ and /ʔ/ alternation would be retrieved, and the entries art, fished and raft would not be returned.

In order to determine the conditions under which alveolar flapping occurs, the queries

O/ P: =DX= and P/ P: =DX=

can be used to retrieve phonological items which contain DX (/ɾ/) and corresponding orthographic items. Such a list might include

ability	/əbílɪɾi/
city	/síɾi/
facility	/fəsílɪɾi/
letter	/lέɾə/
petty	/pέɾi/
responsibility	/rɛspɑnsɪbílɪɾi/
writing	/ráyɾɪŋ/

Flapping occurs in a descending stress pattern, e.g. city letter, petty, writing in which a stressed vowel precedes the flap and an unstressed vowel follows. In addition, the flap appears to occur between unstressed vowels when the sequence represents the noun affix -ity, as in ability. To check this, the query

P/ O: =ITY, S: N

could be used to retrieve all nouns ending in -ity, and the subset involving affixed forms (i.e. excluding city, pity) could be examined for occurrences of flapping.

B. Word Error Recognition testing

The interactive facility can be used to examine the kinds of word recognition errors which might occur in a speech understanding system due to indeterminacies in segment labelling. If a string is completely specified as /likɪŋ/(coded LIYKIHNX), then it matches a single word, leaking. However, if labelling is less precise, then alternative (and incorrect) word matches might occur. Using the interactive retrieval system, alternative labels and resulting word matches can be examined for any given lexicon.

In the example above, the labelled string might be

L ⟨VOC HIGH ANT⟩ K IH NX

with the stressed vowel represented as a set of features: vocalic, high, anterior. Resulting word matches might include leaking and licking.

If the initial consonant is also specified as a set of features (consonant, sonorant, continuant), as in the string·

⟨CON SON CONT⟩ ⟨VOC HIGH ANT⟩ K IH NX

then the resulting word matches might be leaking, licking, reeking. If the K is specified less precisely as a voiceless stop, word matches might include leaking, licking, reeking, leaping, ripping.

The interactive facility allows the system designer to easily determine the nature of possible incorrect matches due to phonological indeterminacy, especially as the size of the lexison increases.

C. Comparative Linguistic Relationships

If the data base is represented as an orthographic list with two associated phonological lists representing two languages or dialects, the interactive system can be used to discover systematic sound correspondences, and to aid in the study of dialect relationships and historical reconstruction.

A sample data base might be:

Gloss	Language A	Language B
a fish	plaa	pa
to have	mii	mia
no, not	plaaw	paw
brother	phii	fia
bamboo	phaay	fay

The query

B/ A: PL=

would retrieve those items in language B which correspond to items in language A with initial /pl-/ clusters, e.g. pa and paw, indicating that consonant cluster simplification may have occurred in language B. The query

B/ A: =IYIY

would retrieve those items in language B which correspond to items in language A with final /-ii/, e.g. the diphthongized mia and fia.

A large data base could be accessed in this way to discover systematic correspondences between languages A and B, such as the correspondences /pl-/:/p-/, /m/:/m/, /ph-/:/f/, /-ii/:/-ia/, /-aa/:/-a/, etc.

The flexibility of the interactive system, combined with the linguistic intuition of the user, can be used to specify and retrieve any set of correspondences, without the need to format the data according to initial consonants or clusters, vowel nuclei, finals, etc. Information such as tonal contours and stress can also be represented and accessed.

REFERENCES

[1] Retz, D. L., J. R. Miller, J. L. McClurg, B. W. Schafer, Elf Kernel Programmer's Guide, Speech Communications Research Laboratory, Santa Barbara, California. April, 1975.

[2] Feldman, J. A. and P. Rovner, "An ALGOL-based Associative Language," Comm. ACM, Volume 12, August, 1969, 439-449.

[3] Barnett, J. A., A Phonological Rules System, TM-5478/000/00, System Development Corporation, Santa Monica, California, 1975.

[4] Bobrow, D. G. and J. B. Fraser, "A Phonological Rule Tester," Comm. ACM, Volume 11, November, 1968, 766-772.

[5] Friedman, J. and Y. C. Morin, Phonological Grammar Tester:

Description, Natural Language Studies No. 9, Phonetics Laboratory, The University of Michigan, 1971.

ACKNOWLEDGEMENT

This research was supported in part by the Advanced Research Projects Agency of the Department of Defense through Contract N00014-73-C-0221 administered by the Office of Naval Research Information Systems Program.