A MORPHOLOGICAL, SYNTACTIC AND SEMANTIC SEARCH ENGINE FOR HEBREW TEXTS.

UZZI ORNAN

Visiting Professor, Computer Science, Technion – I.I.T.
Scientific Director, Multitext, Multidimensional Publishing Systems
ornan@cs.technion.ac.il, uornan@actcom.net.il

Abstract

This article describes the construction of a morphological, syntactic and semantic analyzer to operate a high-grade search engine for Hebrew texts. A good search engine must be *complete* and a*ccurate*. In Hebrew or Arabic script most of the vowels are not written, many particles are attached to the word without space, a double consonant is written with one letter, and some letters signify both vowels and consonants. Thus, almost every string of characters may designate many words (the average in Hebrew is almost three words). As a consequence, deciphering a word necessitates reading the whole sentence. Our model is Fillmore's framework of an expression with a verb as its center. The engine eliminates readings of words unsuited to the syntax or the semantic structure of the sentence. In every verbal entry of our conceptual dictionary the features of the noun phrases (NP's) required by the verb are included. When all the correct readings of all the strings of characters in the sentence have been identified, the program chooses the proper occurrences of the searched word in the text. Approximately 95% of the results by our search engine match those in the query.

1.Introduction

It is easy to construct a search engine that, in a given text, will find all the occurrences of the string of characters specified in the query. In Hebrew script, however, the string of characters that makes up a word may also be interpreted as designating other words. Almost every word in Hebrew script can be read as one of an average of three words. This is because Hebrew script is fundamentally defective: (1) Most vowels in a given word have no sign in the script. (2) Particles are attached with no intervening space to the string of characters that makes up the following word. (3) A geminated consonant is written as one letter, like a not-geminated consonant. (4) Several letters serve as both vowels and consonants. Threfore, it is impossible to identify the word stated in the query by its form: if we try to do so, we would obtain all the occurrences which are written in the same way but are, in fact, different words. Since only 20-30% of the words so obtained are actually occurrences of the required word, the users have to check every word in the result obtained in order to decide whether it is actually the one they want.[1] In order to solve this problem, some systems recommend that every query should contain some other words that are often found close to the stipulated word.[2] But such a search may lead to a loss of important occurrences of the required word. Neither a frequency list of words nor another statistical device can be an ultimate answer in our search of accurate and full device. A statistical approach ensures that some mistakes or

[1] This happened in the case of the programs of Pador, Taqdin, Dinim and others, who offer a search engine for legal texts. It is well known that many lawyers have stopped using them, and prefer to conduct a manual search.

[2] For example, the Contahal company suggested conducting a cross-check.

omissions will always exist. Also, eliminating certain readings by an examination of the words in the short context will not ensure *completeness,* nor will it ensure *accuracy*, since a large number of the strings that appear in the result will not be relevant to the question. (Choueka and Lusignan, 1985; Choueka, 1990). We can obtain a correct reading of a word only if we can make a correct reading of the whole sentence. In order to do this, we must eliminate all the unsuitable readings of every string of characters in the sentence, and leave only one reading. To this end, we had to go through the following stages:

1. First, we adopted a phonemic script, a method of writing Hebrew in Latin characters, in which each vowel has its character, the particles are separated from the following word, geminated consonants are represented by two identical letters, and vowels and consonants are given completely distinct letters .[3]
2. Now we are able to carry out a morphological analysis revealing all the word's components. By examining the results, the correct reading could be clearly seen. This would be impossible in Hebrew script. We constructed a complete, exact morphological analyzer for Hebrew words, which also identifies inflections and attached particles.
3. Having perfected the morphological analyzer, which provides a complete set of details for the analysis of any possible reading of a string of characters, we could write a program which checks every suggested reading of a word, and eliminates readings unsuitable to the syntax of the possibly required sentence.
4. Even a syntactic reading does not ensure that each of the strings in the sentence is indeed a proper reading of the relevant word. Syntactic elimination may leave many words that do not suit a meaningful sentence. Further semantic eliminating is required.
5. For this purpose we compiled a complete conceptual dictionary of the Hebrew language. It is based on Fillmore's ideas about case grammar (Fillmore, 1968), according to which the verb is the center of the expression: it is a function whose arguments are the noun phrases. In every conceptual entry in our dictionary of verbs there appear the semantic, syntactic and morphological features demanded by the verb to exist in the NP's of the sentence, -- including the prepositions, which precede them. Since the dictionary includes also the features of the arguments (NP's) in the sentence, it eliminates readings of words that are suitable syntactically but not semantically. Semantic check enables us to discriminate both between different readings of same string of Hebrew characters as well as between the different meanings of each of the readings.

In this way we completed the necessary basis for the production of an excellent search engine: it will respond to any question only with the occurrences which bear the stipulated meaning, even though the same reading of the characters may have several meanings. The contents of the article are as follows:

In section 2 we shall explain how we establish all possible readings of a string of characters. Section 3 shows how we use syntactic features to eliminate readings that do not fit the syntactic context; then we describe our conceptual dictionary. Section 4 shows how we can eliminate readings that are possible syntactically but not semantically. Finally, in section 5 we

[3] The phonemic script contains four diacritics: ××, ¬, ¡, Ð, also À (or `) and ¿ (or ‘). See ISO-259-3 (available in http://www.cs.technion.ac.il/~ornan/ papers). Several programs for analysis of Hebrew morphology use the regular Hebrew script also for the output. See Bentor et al 1992, Segal, 1999, Carmel and Maarek, 1999 (a statistical filter based on Bentor et al). The problem is that in this way one can show the diverse readings only with the traditional Hebrew dots and points, many of them superfluous. Our method is clearer since we use Latin characters through the whole work.

shall explain how we choose the appropriate meaning of the word by using the dictionary. Section 6 concludes the article.

2. The morphological stage

Our algorithm consists of three stages: morphological, syntactic and semantic. Here we shall describe the first stage, the morphological. The strings of characters are taken from the Hebrew text in Hebrew script, and every string is analyzed. As was mentioned above, Hebrew script containsonly some of the vowels[4] attaches particles to the following word, and does not use double characters to specify geminated letters (see Ornan, 1991); also, some of the characters serve either as vowels or as consonants. It is advisable to be able to read the text in a script that does not have these disadvantages[51]. We use the phonemic script of ISO (FDIS 259-3). Thus, for instance, the Hebrew word HRKBT can be read in any of the following ways:

hirkabta, hirkabt, harkabat
ha-rakkebt, h-rakabt, h-rakabta

In the morphological stage, each of these possibilities is written at the beginning of a separate line, followed by all the grammatical details of the reading:

```
hrkbt   V  hirkib    ,-,-,ta ,p,2,+,#,s -,-,-,-,-
hrkbt   V  hirkib    ,-,-,t  ,p,2,#,+,s  -,-,-,-,-
hrkbt   N  harkaba   ,c,-,t  ,-,3,#,+,s  -,-,-,-,-
hrkbt   N  rakkebt   ,a,-,-  ,-,3,#,+,s  -,-,-,-,- ha-
hrkbt   V  rakab     ,-,-,t  ,p,-,#,+,s  -,-,-,-,- h-
hrkbt   V  rakab     ,-,-,ta ,p,-,+,#,s  -,-,-,-,- h-
lq×nwh V  laqax     ,-,-,nu ,p,1,+,+,p  3,#,+,s,h
```

The given Hebrew word is the first column. The second column is the category. The third column is the lexical entry. The following column gives the status of the word (construct, inflected or absolute). Next come the prefix and suffix of the word, the tense (for a verb), person, gender (masc., fem. or both) and number (s or p), and then details about person, gender and number of the attached pronoun (see the last example *lq×nwh*), and the attached pronoun itself. The last column specifies attached particles.

This morphological analysis is based on a program which uses a complete lexicon[6], based on a comprehensive grammar of all possible Hebrew word-patterns – including, of course, all inflections, regular and irregular.

3. The syntactic stage

Each of these lines presents one possible reading of the given word. But usually only one reading is acceptable in any given sentence. Therefore, we must eliminate those readings, which are morphologically correct, but incorrect in the given context.[7] The first elimination is syntactic, and it is done in the realm of one "Syntactic Unit", i.e., a clause which includes one verb and is bounded by a "sign of separation", mainly subordinating or certain coordinating particles.[8] At this stage all possible analyses of the strings of characters are displayed. Now, the program attempts to combine each line of every word with every one of the lines of all other words. The correctness of the combination is checked with all possible

[4] Ide and Vèronis (1998:2) mention this as a cause of the need to "disambiguate" Semitic languages. I would have been more correct for them to refer in their article to the system of writing rather then the languages.

[5] See Ornan, 1987, Ornan and Katz, 1994. See note 3.

[6] In general, we used Even-Shoshan, 1994, which is still considered to be the best.

[7] An interesting attempt to decipher a text in Hebrew script - Nirenburg and Ben Asher, 1984.

[8] "Short context rules" are not a satisfactory solution, while our full syntactic analysis is easily derived from the Conceptual Dictionary, as explained in what follows.

sequences of other words. Practically, only a small number of these combinations make a sentence that is syntactically correct.[9] How is the syntactic test performed?

The program computes every combination of possible strings of words. For example, giving the sentence HBWQR ZRªH £M£ ªMH (in Hebrew script – "hot sun rose this morning") will render the following analysis of all readings of the words of this sentence:

```
hbwqr N boqr       ,a,-,-   ,-,3,+,#,s  -,-,-,-,- ha-
hbwqr N boqer      ,a,-,-   ,-,3,+,#,s  -,-,-,-,- ha-
zr×h   N zar×a     ,a,-,-   ,-,3,#,+,s   -,-,-,-,-
zr×h   N zer×      ,i,-,-   ,-,3,+,#,s  3,#,+,s,h
zr×h   V zara×     ,-,-,h   ,p,3,#,+,s -,-,-,-,-
¡m¡    N ¡amma¡ ,a,-,-   ,-,3,+,#,s  -,-,-,-,-
¡m¡    N ¡amma¡ ,c,-,-   ,-,3,+,#,s -,-,-,-,-
¡m¡    N ¡em¡   ,a,-,-   ,-,3,#,+,s -,-,-,-,-
¡m¡    N ¡em¡   ,c,-,-   ,-,3,#,+,s -,-,-,-,-
¡m¡    A ¡amma¡ ,a,-,-   ,-,-,+,#,s  -,-,-,-,-
¡m¡    V ¡imme¡ ,-,-,-   ,i,2,+,#,s  -,-,-,-,-
¡m¡    V ma¡    ,-,-,-   ,p,3,+,#,s -,-,-,-,- ¡e-
¡m¡    V ma¡    ,-,-,-   ,r,-,+,#,s  -,-,-,-,- ¡e-
×mh    N ×ema   ,a,-,-   ,-,3,#,+,s -,-,-,-,-
×mh    A ×amma  ,a,-,-   ,-,-,#,+,s -,-,-,-,-
×mh    N ×amma  ,a,-,-   ,-,-,#,+,s -,-,-,-,-
```

This short expression provides 144 sentences to be checked: 2x3x8x3=144. The syntactic stage will eliminate the great majority of invalid sequences of possible readings. We shall not discuss them all here – only make some remarks about a few clear cases for elimination. For example, the readings *boqr, boqer* cannot function syntactically as the subject of the sentence, since they are masculine, and no verb in the rest of the sentence which is not preceded by subordinating ¡*e*- ("that") is masculine (agreement is needed). Similarly, the second word cannot be *zar×a*: a feminine noun, and no verb agrees with *zar×a* (as subject) in the analyses of the other words.

First, the program looks for a verb. When a verb is identified, the program checks possible nouns that can be the syntactic subject. It then checks other NP's and PP's, possible adjectives and adverbs. Mainly because the order of words in Hebrew is rather free, the syntactic stage usually leaves a few possible sentences that may be accepted as proper readings of the input sentence from the syntactic point of view. But some of these possibly correct syntactic readings may possess improper semantic!! features, which should not be accepted.

We have a special treatment for sentences without a verb (this may occur in Hebrew and other languages, especially Semitic): if the program does not identify a verb in the input sentence, it adds the verb *haya* ("to be") in the appropriate gender, number and person, and the review process is repeated. Our dictionary of verbs is described below. Here we may remark that the verb *haya* appears in more than one lexical entry. One of them should be accepted. We shall preface the description of the stage of semantic elimination with an account of its fundamental characteristics.

4. The conceptual approach

4.1 **Introduction**

Every natural language is a means of describing the world. It contains symbols of concepts (concrete, abstract or imaginary). Speakers of the language use these symbols in order to designate these concepts as they occur in the world.[10]

[9] There has been much research on syntactic analysis by means of a computer program; for instance, Wintner and Ornan, 1995. Herz and Rimon, 1992, also deal mainly with syntactic problems. Levinger et al., 1995 demonstrate methods of eliminating syntactically incorrect morphological readings. See, too, Levinger, 1992.

[10] See Ostler, 1995:221, who emphasizes the world-outlook common to all languages:
"… there is a fair degree of comparability among the units engaged by each language, just as there is a fair degree of similarity between the features of the human condition that they describe. We all have the same sense-organs, live in modern Western societies with other human beings, confront the same tasks of providing food, clothing and shelters for

It is true that most of the words in every natural language are symbols of concepts, of actions, and of the relationships between them. But, as was pointed out above, every natural language also contains other, organizing elements, which do not symbolize concepts or actions and do not refer to the extra-linguistic world. These elements organize the other words around them: this is the difference between organizing elements and symbolic terms. By "Organizing Elements" we are not referring only to what are called "grammatical words", such as *ki* in Hebrew, or "that" in English – words which do not refer to any entity in the world outside the language, but give information about the other words in the expression; these words (such as *ki*, "that", ¡*ello*, "whose") inform us, for instance, that what follows them is intended to provide details of whatever preceded them, or to describe it in a particular way. "Organizing Elements" also include morphological details which have a linguistic meaning, such as indications of gender (*bianco – bianca* in Spanish), of number, (*boy – boys* in English), or person (*vide – videsti* in Italian) a hint to the definiteness of what follows (*a – the* in English), case endings which indicate the syntactic function of the concept symbolized by a noun in relation to an operation in the world indicated by a verb in the expression (in Arabic, *baytuun* as subject – *baytaan* as object), and so forth. All of these are morphological means, which serve to organize conceptual symbols.

In contrast to the conceptual elements, the organizing elements in the expression differ as between languages not only in their external form, but also in their nature. Languages differ from each other in their systems of organizing symbols. Thus, what is unique in every natural language is concentrated in the organizing elements, and far less in the lexical sphere. The dictionary that we constructed is based on these assumptions. It is first and foremost a "dictionary of human concepts"; but we also had to include the organizing elements in it. We shall now describe this dictionary.

4.2 The two parts of the conceptual dictionary

Standard dictionaries are arranged in alphabetical order, with the category of each lexical entry noted. Our dictionary is divided from the first into two main dictionaries: one for nouns, and one for verbs. This will shortly be discussed in detail; but, first, we may observe that since the same Hebrew word frequently serves to symbolize several concepts (whether this be a homograph, polysemy or homonym), we add an index number to the lexical entry: for instance, c*ir1* ("delegate") is a different concept from c*ir2* ("hinge, pivot"), even though in Hebrew both of these concepts are symbolized by the same word, c*ir*. Similarly, ¡*eq* '*1* means low barometric pressure, whereas ¡*eq* '*2* means an electrical wall-plug: both concepts are symbolized by ¡*eq* '. In both of these instances, we introduced two different entries.

4.3 Dictionary of noun concepts

An entry in the dictionary of noun concepts consists of a list of the essential features of the concept. Here are some examples: The conceptual entries of *bayt* read as follows:

bayt1: {construct} {site} {receptacle}
{concrete} {property}.

bayt2: {site} {receptacle} {intimate}
{family}.

bayt3: {word} {information}
{work of art} {poetry}.

The words in curly brackets indicate features of the concept (in our dictionary

---

ourselves and our families, are confronted or aided by much the same degree of technical progress and so on." See also Whorf 1956:138ff. Both authors speak of Western societies.

they are in Hebrew, but have been translated for this article). We began the work with an arbitrary list of about 130 features of concepts, but eventually more were added in the course of work in order to define new concepts, and we now have about 170 features.[11] The reader will see that the concept *bayt1* refers to the English word "house", *bayt2* to "home", and *bayt3* to "stanza".

The idea of a conceptual dictionary was conceived as a means of constructing an infrastructure for comprehensive processing of the Hebrew language, and not only for the construction of an efficient search engine. This base has already served in the construction of a Hebrew "Reading Machine" for the blind[12]. Recall that in order to read a Hebrew text the whole sentence must be read. Sometimes a shorter context is sufficient. The conceptual dictionary is intended to enable the sentence containing the given word to be read accurately by using a sophisticated procedure that takes into account all the possible readings of every word in the sentence and by reading the whole sentence, and not simply word by word. We must now describe the dictionary of verbs.

4.3 The dictionary of verbs

The dictionary of noun concepts by itself cannot activate the algorithm required for correct reading of the Hebrew sentence. A dictionary of verbal concepts is also required.[13] C.C. Fillmore (1968) opened new linguistic horizons by putting the verb in the center of the expression, and showing how all the other parts of the sentence should obey the demands of the verb. (Tesnière should be mentioned here as the "father–figure", as Somers 1987, p.1 emphasizes, but note what follows on the same page, as well as in Ch.2.) We exploit this concept to the full, and extend it to build a dictionary of conceptual entries related to actions in the world.[14]

Therefore, the dictionary of verbs contains in the entry of every single verb everything that that particular verb requires to be included in the sentence. First of all, the verb's entry includes the answer to the question: what specific thematic functions are required in the sentence to which this particular verb is central,[15] and what semantic features must the noun phrases which perform these thematic functions possess.[16] [2] The thematic functions themselves are common to all men: for example, the thematic function "agent" or "experiencer" exists in most sentences in various languages. This also applies to the thematic functions "theme" and "instrument".

In our dictionary, however, the verbal entry also contains organizing elements: in the first instance, the prepositions that the verb requires or allows to be placed before the noun phrases. We included the prepositions in the verb dictionary in order to solve the problem of the prepositions individually and rigorously. Basically, a preposition is an organizing element: it springs not from reality, but from the conventions appropriate to each particular language, and relates to reality only partly and, in general, quite vaguely. In

[11] Miller's WordNet is a mine of features, many of them but not all have been used in our engine. See especially Miller, 1993.

[12] In the years 1996-98 an Israeli company (Eastek) developed a "reading machine" for the use of the blind in Israel, using this base in its first version.

[13] Stern's *Verb Dictionary,* 1994, is not a conceptual dictionary. Although it includes in every lexical entry the particles to be found in expressions in which the verbal entry is central, it does not relate to thematic arguments and their semantic features.

[14] In honor of Fillmore my students call our conceptual dictionary "Fillmore Dictionary".

[15] The same ides is sometimes called, less clearly, "selectional restrictions" (Chomsky, 1965,1984).

[16] Compare some examples of entries, or "case frames", suggested by various authors in computational linguistics in Somers, 1987 illuminating book, mainly in Ch.11.

organizing the material in this way, there is no need to deal with the problem of classifying "types" of verbs (see the discussion in Somers 1987: 70-74, and 283 et seq.), or to categorize them according to "selectional restrictions" (see note 15).

Secondly, some adverbs must appear in the verb's entry in the dictionary: some as optional elements; others, occasionally, as necessary elements. In every case they are labeled as fulfilling a secondary thematic function, to which we give the variable code NP3 or NP4 (NP1 agrees in person, number and gender to the verb, i.e., it is the subject, NP2 is usually the theme). Round brackets show that the occurrence of this element is optional. But there are also "external" adverbs, which cannot be included in the lexical entry of the verb even though they may appear frequently in many input sentences. In the main, they indicate the time or place of the action, or function as "sentential adverbs" which describe the external circumstances of the event. When such an adverb appears in an input sentence (it may appear as an unidentified element in a sentence for analysis) we give it the symbol of a noun phrase (NP), with a special index number: NP8 or NP9. In various systems of linguistic analysis it is usual to mark adverbs PP. However, this symbol seems superfluous and we have preferred to mark this element as NP, with a preposition preceding it (in our system, a preposition may also be " "). A similar problem may arise with adjectives. They, too, are not included among the requirements of the verbs but, of course, they occur in the input sentences. We built a separate dictionary for adjectives. It constitutes a separate section of the noun dictionary, and its entries may be added to the noun (with the organizational and semantic limitations of their lexical entries) or function independently as separate NP's.

Thirdly, in many conceptual entries extra details are sometimes included, for example, special modifications of gender or number, the addition of epithets appropriate to one of the required noun combinations, idioms, etc. Another example of elements that are not connected with a concept arising from reality is to be found in verbs expressing thought or speech. They require a subordinate clause which is characterized by the feature "{contents}", but the content is unrestricted (in Arabic grammar these are known as "verbs of the heart"). This is a syntactic condition, and may, therefore, differ among languages. As can be seen, the conceptual dictionary is not purely semantic. It also contains many syntactic elements, and even some morphological details. In our dictionary we have formulated approximately ten thousand verb entries. The noun dictionary at the moment consists of thirty five thousand entries.

Here are some examples of conceptual verb entries as they appear in our dictionary, followed by some explanatory notes.

```
Pitte×1            [= develop a film]
   NP1 AGENT  {human, role, org}
   NP2  THEME  "et" {printed matter,
                          picture}

pitte×2            [= open a knot]
   NP1 AGENT  {human, role, org}
   NP2  THEME  "et" {knot, "×gor"}

pitte×3[173]       [= carve]
   NP1 AGENT  {human, role}
   NP2 THEME  "et " {writing, word,
                          picture }
  (NP3) INSTRUMENT "b-" {tool,
                       sharp, acute}
  (NP4) INSTRUMENT  "'al" {solid,
                    platform, article}
```

As can be seen, the Hebrew verb *pitte*× signifies three conceptual entries, which in other languages could well be expressed by three completely different and dissimilar words.

In the following examples there are also limitations of organizing elements. Here we should add some remarks about special symbols and explanations of the names of the thematic functions: % indicates that the succeeding word signifies a syntactic structure, and not a thematic function. Double quotation marks

[17] There are some other meanings to this verb, but in order to explain our approach these three suffice.

(“ ”) indicate Hebrew words, particularly prepositions (such as “et”, “b-”, “ ¿al”) which are required by the verb or are parts of idioms. Round brackets “( )” denote an optional function which is not obligatory to the sentence to be analyzed. A diagonal stroke “/” indicates another possibility, shown in the expression which follows it.

```
hebin          [= understand]
  NP1 EXPERIENCER {human, role,
                              org}
  NP2  "¡e-" %SENTENCE
 / NP2 “et” AIMED-AT {abstract,
                              info}

¡imme¡1      [= serve ]
  NP1  INFLUENCER {-}

  NP2  THEME  {human, org}
  (NP3) {"l-"/"k-"} GOAL {action}


¡imme¡2      [= be used as ]
  NP1 THEME {human, instrument,
                site, construction}
  NP2 "btor"  FUNCTION { human,
        instrument, site, construction}

‘arak1          [= set a table]
  NP1 AGENT {human, role}
  NP2 "et" THEME {"¡ulxan"}[=table,
                          an idiom]

 ‘arak2          [= organize ]
  NP1 AGENT {human, org.}
  NP2  "et" THEME {act, happening}

 ‘arak3          [= edit ]
  NP1 AGENT {human, org.}
  NP2  "et" THEME {printed_matter,
                       work of art}

zarax           [= rise (sun) ]
  NP1   THEME  {source_of_light,
            source_of_heat, strong}

rakab           [= ride ]
  NP1 AGENT {human}
 (NP2) "‘al" THEME {four_legged_
                  animal, vehicle}
 (NP3) "l-" TO-LOC {site, place,
                happening, human}
 (NP4) "mi-" FROM- LOC {site,
          place, happening, human}
```

The Hebrew preposition *et* is sometimes rendered as " " (especially when it precedes an entry noun without a definite article); but, since this is always so, it is not worth printing " " separately for each entry; so this possibility is included in the program.

Another example: the expression *maca` xenn* is an idiom, and it seems as if one could treat the combination as one word. But since there could also be an expression in which the two words of the idiom were separated, and in view of the morphological difficulties to which such a solution could lead, we prefer to formulate it as a simple one-word verb, located in a special lexical entry *maca`1*. This verb requires obligatory completion in this expression – *“xenn”*, with NP2 status.

```
maca`1          ["NP3 likes NP1"]
  NP1  AIMED-AT  { }
  NP2   THEME  {"xenn"}
  NP3  "b-‘einei" AGENT {human,
                       animate}
```

After our description of the conceptual dictionary, we shall now describe the process of semantic elimination.

5. The semantic stage

It will be recalled that we first identified a verb among the readings of the words, and then dealt with the elimination of syntactically improper readings. We now turn to the conceptual dictionaries to see whether the NP's accord with the expected thematic role of the proposed verb, and whether the NP’s contain the appropriate semantic features. This procedure is repeated for each possible reading of a string as a verb in order to discover all possible interpretations of the sentence. The final results usually contain one interpretation only, the intended one. But sometimes more than one interpretation is received. This is for one of two reasons: either the program discovers a true and appropriate reading that a human being did not think of, or the interpretation does not fit conditions in the real world. To include some means of checking knowledge of the world in the program would, of course, be a formidable problem. But these results are quite rare, and presented in the results only as another possibility, besides the correct one.

6. Conclusion

The analysis and identification of correct readings of words in Hebrew script is far from being a simple task. The correct reading of any word is achieved only as a result of reading the rest of the words of the complete clause and sentence. We had
to invest much work to solve this problem and to build a complex system of programs before we could have achieved a high-grade search engine, which is far better than other existing suggestions. See appendix for comparison.

As noted above we succeeded in producing this engine only on the assumption that a Hebrew word must be read not on its own but in accordance with the reading of a complete sentence. We found that we achieved a powerful program, constituting a comprehensive infrastructure, for processing the Hebrew language for the computer. This infrastructure has already produced other results (the reading machine for the blind, see note 12), and it enables us to begin to work on automatic translation from Hebrew to other languages — a task which has never yet been attempted.[184] It seems that this method of translation by computer may be suitable to any language, and could be a contribution to translations from other languages.

7. References

Bentor,E., A.Angel, D.Ben-Ari-Segev and A.Lavie. 1992. Computerized Analysis of Hebrew Words in *Hebrew Computational Linguistics*, ed. by U. Ornan, G.Arieli and E.Doron, Ministry of Science and Technology, pages 36-38.(Hebrew).

Carmel, David and Yoelle Maarek. 1999. Morphological disambiguation for Hebrew. In *Proceedings of the 4th International Workshop NGIT-99, Lecture notes in computer science 1649.* Springer Verlag, pages 312-325.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Chomsky, N. *Lectures on Government and Binding*.. .Foris Pub.

Choueka, Y. and Serge Lusignan.1985 . Disambiguation by Short Contexts. In *Computers And Humanities* Vol 19:147-157.

Choueka Yaacov.1990. Responsa: An Operational Full-Text Retrieval System. In *Computers in Literary and Linguistic Research*. edited by J .Hamesse and A. Zampoli. Champion-Slatkine Paris-Geneve . Pages 94-102

Even-Shoshan, Avraham.1987. *The New Dictionary*. (Hebrew).

Fillmore,C.C.1968. The Case for Case. In *Universals in Linguistic Theory* . Edited By E.Bach and R.Harmes. Holt, Rinehart and Winston, New York, Academic Press. Pages 59-81.

Herz,Y. and M.Rimon.1992. Diminishing Ambiguity by Short-Context Automaton. In *Hebrew Computational Linguistics* .Edited by U. Ornan, G.Arieli and E. Doron. Ministry of Science and Technology. Pages 74-87.(Hebrew).

Ide, Nancy and Jean Véronis.1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Compuational Linguistics*24: 1-40 .

ISO 259-3.1999 .*Conversion of Hebrew Characters Into Latin Characters. Part 3: Phonemic Conversion*. ISO/TC46/SC2.

Levinger, Moshe. 1992. Morphological Disambiguation in Hebrew. Research Thesis for MSc in Computer Science. Technion. Haifa (Hebrew).

Levinger,Moshe, Uzzi Ornan and Itai Alon. 1995. Learning Morpho-Lexical Probabilities from an Untagged Corpus with an Application to Hebrew. *Computational Linguistics* 21: 383-404 .

Miller, George A.1993. *Nouns in WordNet.* Web file.

Nirenburg, Sergei and Y. Ben Asher. 1984. HUHU : Hebrew University Hebrew Understander. *Journal of Computational Ling*uistics Vol.9:161-182.

Ornan, Uzzi. 1987. Hebrew Text Processing Based on Unambiguous Script. .*Mishpatim* 17:15-24.(Hebrew)

Ornan, Uzzi.1991. Theoretical Gemination in Israeli Hebrew. *Semitic Studies in honor of Wolf Lwslau.* Edited by Alan S, Kaye. Otto Harrassowitz, Weisbaden. Pages

[18] We are not referring to computerized bilingual dictionaries, but to a full translation.

1158-1168.
Ornan, Uzzi and Michael Katz. 1994. A New Program for Hebrew Index Based on the Phonemic Script .*TR #LCL 94-7* (revised). Technion - I.I.T.
Ostler, Nicholas. 1995. Perception Vocabulary in five Languages – Towards an Analysis Using Frame Elements. In Steffens Petra (editor) *Machine Translation and the Lexicon*. Springer Verlag. Pages 219-23.
Segal, Erel,1999. Hebrew Morphological Analyzer for Hebrew undotted Analysis. Thesis for MSc in Computer Science. Technion. Haifa (Hebrew).
Somers, H.L. 1987.*Valency and Case in Computational Linguistics.* Edinburg University Press.
Stern, Naftali. 1994. *The Verb Dictionary.* Bar-Ilan University. (Hebrew).
Wintner, Shuly and Uzzi Ornan. 1995. Syntactic Analysis of Hebrew Sentence. *Natural Lnaguage Engineering* 1:261-288.
Whorf, Benjamin Lee. 1956. The Relation of Habitual Thought and Behavior to Language. In Leslie Spier (editor) *Language, Culture and Reality* ,essays in memory of Edward Sapir.1941. Pages 75-93. Reprinted in John B. Carrol (editor) *Language, Thought and Reality*. M.I.T. Press. Pages 134-159

.

.