Reference:

LEFFA, Vilson Jose. Clause Processing In Complex Sentences. In: FIRST INTERNATIONAL
CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, Proceedings of the First International Conference on Language Resources and Evaluation. Granada, Espanha: 1998. v. 2, p. 937-943.

Clause Processing in Complex Sentences

Vilson J. Leffa

Universidade Católica de Pelotas

Rua Felix da Cunha, 412

96010-000 - Pelotas, RS, BRAZIL

[leffa@via-rs.net]

Abstract

The purpose of this investigation is to propose and test an algorithm for the segmentation of complex sentences into clauses. The algorithm is built after the parts of speech for each lexical item are assigned. Formal indicators of subordination and coordination, along with information about the valence of the verbs found in the immediate context are used to mark the beginning and end of each clause. When the clauses are identified they are classified into either a noun or an adverb, using information provided by the surrounding context. The algorithm was tested using a machine translation system developed by the author, which included an English/Portuguese dictionary, a part of speech tagging system and the ability to introduce rules, including the ones required by the algorithm. The results showed that out of 1659 clauses, randomly taken from a 10,000,000-word corpus, more than 98% were correctly segmented and 95% correctly classified into nouns or adverbs.

The major problems found in segmenting and classifying the clauses included conjunction ambiguity, verbs that belonged to more than one subcategorization, and the sharing of the same subject by different clauses.

Introduction

An indispensable task in Natural Language Processing (NLP), regardless of the linguistic approach favored by any system, is the identification of the structure that underlies the sentence. This task involves the ability to partition a given sentence into hierarchical segments, not only at word level but also above the word (such as phrases and clauses), and below it (including prefixes and suffixes).

Segmentation at all these levels is important for NLP, but some areas have received very little attention. One area that has been particularly avoided by researchers is segmentation of complex sentences into clauses — such as the segments separated by "/" in example (1). It is impossible, however, to process a complex sentence if its clauses are not properly identified and classified according to their syntactic function in the sentence (subject, object, etc.).

(1)*That the girl refused the flowers / surprised the boy / who was trying to be nice.*

This investigation proposes to approach the problem of clause resolution by exploiting one of its basic properties — that is, the ability of clauses to be ultimately reduced to a noun, an adjective, or an adverb, no matter how long they are, or how many other clauses they may have embedded in them. This reduction into a part of speech category is necessary to produce well-formed sentences, since clauses ultimately behave as if they were one word. In a sentence like (2), for example, the verb *surprises* can only be correctly inflected if the subordinate *clause That they refused the flowers* is properly identified as a noun and assigned the function of subject. This can only be done if the subordinate clause is previously segmented at the right places and processed before the main clause.

(2) *That they refused the flowers surprises me*.

Towards an Eclectic Approach

The only study in NLP, to my knowledge, that has explicitly addressed the issue of sentence segmentation into clauses is Stefanini (1993), who, in her multi-agent system for natural language treatment, designates one agent for this task (p. 150 and ff.). This segmenting agent knows the formal indicators of subordination and coordination and should be able to build trees that represent the structure of complex sentences. The author, however, leaves many problems unsolved, including the inability to distinguish between some verb phrases and clauses (e.g. *I can work* versus *I want to work*) and the difficulty that arises when clauses are inverted (*If you study the books will help you*), where it is not clear how the system would separate the clauses. The use of finite verbs as the sole criterion for defining a clause should also lead to some problems in the processing of a sentence like (3), where the first clause *Visiting many cities* would not be identified, thus making it impossible to locate the right subject for the verb *make* — a problem in many NLP systems, which would incorrectly assign *cities* as the subject, thus producing (4).

(3) *Visiting many cities makes me tired.*

(4) **Visiting many cities make me tired.*

The scarcity of studies on the specific topic of clause segmentation forces us to look at the problem from a broader linguistic perspective, where we find that the complex sentence has raised much more interest in pedagogical grammars than in purely theoretical approaches. While Quirk et al. (1985), for example, devoted more than 200 pages to the topic in their *Comprehensive grammar of the English language*, references to clause segmentation in the theoretical literature are scarce and incomplete, scattered over a wide range of topics such as the Accessibility Hierarchy for relative clauses in Keenan and Comrie (1977), the use of nominalizations for determining theme in functional grammar in Halliday (1985), the Inflection Phrase in X-bar syntax in Chomsky (1986), the rules for anaphora in the binding theory (Goodluck, 1991).

In terms of sentence partitioning, a review of the literature suggests three ways in which a sentence can be segmented to the clause level: (1) starting with the first word in the sentence and processing it from left to right, word by word, until all the clauses are identified; (2) starting with formal indicators of subordination and coordination and proceeding

until the end of the clause is found; (3) starting with the verb phrase, identifying the verb type and locating its subject and complements.

Word by word processing is one of the most traditional approaches in linguistics and can be attributed to Hockett (1955) and his finite-state grammar. According to Chomsky (1957, p. 20), in his criticism of Hockett's model, the speaker when producing a sentence "begins in the initial state, produces the first word of a sentence, thereby switching into a second state which limits the choice of the second word, etc. Each state through which he passes represents the grammatical restrictions that limit the choice of the next point in the utterance." For Chomsky, finite-state grammars are limited and incapable of accounting for all the possible sentences in a language, including embedded clauses. In example (5), the clause *who refused the flowers* could not be accounted for by the finite-state model. In spite of Chomsky's criticism, finite-state grammars have provided the theoretical background for many different approaches, both in NLP models, including the Augmented Transition Network (ATN), as proposed by Woods (1970), and in human processing models, including listening comprehension, as proposed by Brazil (1995).

(5) *The girls who refused the flowers surprised me.*

The identification of formal indicators of coordination and subordination has the advantage of counting on the surveys already done by traditional grammars (e.g. Quirk et al., 1985). These surveys have also shown that the identification is sometimes complicated by the use of other signals instead of conjunctions such as inversions (*Had she accepted the flowers...*), certain verb forms (*Having accepted the flowers...*) or even no signal at all (*The girl the boy offered the flowers arrived*). It is also easier to partition a sentence when the clauses themselves are not inverted (*Problems show up if you love money*), than when they are (*If you love money problems show up*), in which case, there are at least three possibilities: (6), (7) and (8).

(6) *If you love money / problems show up.*

(7) *If you love money problems / show up.*

(8) *If you love / money problems show up.*

While finite-state grammars and the use of coordinators and

subordinators imply a left-to-right order of processing, an approach centered on the verb implies processing the sentence is both directions, beginning with the verb phrase, which governs the clause, and spreading to the periphery, until the frontiers of the clause are reached. Valence grammar (Borba, 1996) is a typical example of this approach: the verb is not only an essential part of the clause but the governing center from where control is exercised over each of its arguments. A limiting factor for the application of the model is the valence ambiguity of many verbs such as *love* in the examples above, which may act as either transitive or intransitive, thus making it more difficult to segment the sentence.

Finite-state grammars and valence are the main theoretical approaches suggested by the literature to attack the problem of the complex sentence and its segmentation into clauses, while the use of formal indicators may be seen as an application of either or both theories.

The proposal here, in broader terms, is to use both. The assumption is that language processing occurs over such a wide spectrum of different subprocesses that no single approach can account for all the different procedures involved in NLP. It may be the case, for example, that the identification of complex NPs, a necessary subprocess, is better done in a finite state paradigm, while clause identification would be more adequately addressed under a valence perspective.

It is assumed that the use of different approaches also facilitates the distribution of tasks, performed in both an independent and integrated fashion, including the ability to use information from other levels when necessary. The identification of a noun phrase, for example, may sometimes need to use valence information to decide whether two nouns may or may not combine to form a complex noun. In a sentence like (9), the decision to process the words *student* and *satisfaction* as two NPs depends on the number of arguments required by the verb to give. This would not happen in a sentence like (10), where the same words would belong to the same NP.

(9) *He gave the student satisfaction.*

(10) *Student satisfaction is high.*

Methodology

The methodology used in this investigation to segment complex sentences into clauses can be described in three steps: (1) setting the criteria for deciding what makes a clause a clause; (2) establishing the linguistic rules for segmenting the clauses; (3) testing the rules on a corpus.

The criteria

The basic criterion for deciding whether or not a given segment of language constitutes a clause is the presence of a verb phrase, either finite (*was, goes*) or non-finite (*be, gone, going*). A verb phrase can obviously have more than one verb. This happens, for example, when the main verb is preceded by auxiliaries (*operators* in Halliday's terminology). It does not happen when the main verb is preceded by other main verbs. The sentence *They must have been working* has one verb phrase, since *working* is preceded by auxiliaries. On the other hand, the sentence *They want to work* has two verb phrases, since *want* and *work* are both main verbs, belonging, therefore, to two different clauses.

Coupling verb phrases with clauses, in a one-to-one fashion, admittedly leads to some moot points in theoretical linguistics. First, the criterion classifies not only finite constructions as clauses but also all nonfinite constructions such as the segment *you wearing miniskirts* in the sentence *I don't like you wearing miniskirts* — which is seen by some linguists as a simple phrase, not a clause. The main justification for considering it a phrase is that *wearing* has neither time reference not agreement. We found it easier to treat these segments as clauses, however, considering that they have a subject (*you*), a transitive verb (*wearing*), and a direct object (*miniskirt*).

Second, and in the opposite direction, marginal clauses, which do not present a verb phrase, are not regarded as clauses here. The following sentences (11-13), for example, were regarded as single independent clauses, no matter how many words they have:

(11) *If necessary, he will take notes for you.*

(12) *I do not wish to describe his assertions, some of them offensive.*

(13) *She looked at him expectantly, her eyes full of excitement and curiosity.*

One critical factor for both decisions was of a practical nature. English nonfinite constructions with subjects like the segment *you to work* in the sentence *I want you to work*, when translated into Portuguese, becomes finite with both time reference and agreement. Depending on the surrounding context, *you to work* can produce in Portuguese many different translations, marked for agreement and time such as: *que você trabalhe, que vocês trabalhem, que você trabalhasse, que vocês trabalhassem, que você tivesse trabalhado*, etc.

On the other hand, elliptical constructions such as *if necessary* are exactly the same in both English and Portuguese, with no need to recover the verb. This is admittedly an ad hoc decision and is taken only as far as these two languages are considered. It is not advocated that it would work with any other languages.

Another difficult problem concerned the verb phrase. It is not always easy to decide when a sequence of verbs belong to one or more verb phrases, especially in cases where semantic ambiguities are involved. While it is less difficult to separate clauses based on syntactic restrictions, as in *The book he has described Rome* (two verb phrases), it is extremely difficult to separate them when the decision has to be based on semantic constraints. This can be seen by comparing the two sentences below (14-15):

(14) *The professor is teaching. (one verb phrase)*

(15) *The problem is teaching. (two verb phrases)*

In the first sentence (*The professor is teaching*) there is only one verb phrase and one clause, implying, correctly, that it is the professor who does the teaching. In the second sentence (*The problem is teaching*) there are two verb phrases (*is* and *teaching*) and, consequently, two clauses. Translating the sentence as a single clause would produce a semantically anomalous sentence in a language such as Portuguese, producing something like *O problema está ensinand*o, which would mean that that it is the problem which does the teaching. The difference between treating the sentence as either one clause or two clauses can be seen in the Portuguese translations below (16-17):

(16) **O problema está ensinando.*(one clause)

(17) *O problema é ensinar*. (two clauses)

In general, the one-verb-phrase-one-clause criterion seemed to work satisfactorily, as long as the clauses were correctly separated. The ability to do that depends on the segmentation rules, which will be discussed on the next section.

The rules

The formulation of the rules for segmenting the sentences into clauses were based on two different sources. The first was the *Comprehensive grammar of the English language* (Quirk et al. (1985, pp.918-1146), which probably offers the most comprehensive inventory of clause types in English; from that inventory, 1,513 complex sentences were taken and analyzed. The second inventory was a set of 1,000 sentences taken at random from newspaper texts. These sentences were manually analyzed and classified not only for the purpose of assessing how different types of clauses are distributed over authentic situations of language use but also for the purpose of detecting which clause boundary signals were used besides conjunctions. Table 1 summarizes the findings in terms of clause types.

Clause Type	%
Independent clauses	15.5
Main clauses	18.7
Coordinate clauses	29.2
Subordinate nominal clauses	14.7
Subordinate relative clauses	11.1
Subordinate adverbial clauses	10.8
Total	100.0

Table 1: Distribution of clause types in newspaper texts.

In terms of clause boundary signals, the purpose was to expand the idea of formal indicators of coordination and subordination to include any trait that might be used to mark clause initiation and termination. This trait could be not only the occurrence of certain word types like subjective pronouns (*he, they*), which can mark clause initiation, but also sequences of word types such as a finite verb followed by a finite verb. It can be seen, for example that the sequence *ordered arrived* in the sentence *The books you ordered arrived* indicates that the two verbs belong to two different clauses and that a segmentation signal should be

placed between them.

The correct identification of NPs is also a crucial point in segmenting the clauses correctly, as was seen above in the *student satisfaction* example. Certain NP sequences, however, seem to have zero probability of occurring in the same clause, and thus, would also indicate that a clause segmentation signal should be placed between them. Thus, although the sequence *the boy the flower* can occur in the same clause (*She gave the boy the flower*), the inverted sequence *the flower the boy* may never be found in one clause. It seemed that an NP with the semantic feature +HUMAN, followed by an NP with the semantic feature -HUMAN, would indicate that the NPs belonged to two different clauses. The following examples should illustrate when the NP sequence is and is not allowed (18-19):

(18) **She gave the flower the boy.*

(19) *She gave the boy the flower*.

although, obviously the sequence *the flower the boy* can occur when the two NPs belong to different clauses, as demonstrated in the example below (20):

(20) *She gave the flower the boy wanted.*

The purpose in using the inventory presented in the *Comprehensive Grammar* and the 1,000 sentences in the newspaper texts was to detect these clause segmentation signals, whether explicitly actualized in the text such as conjunctions, or implicitly understood such as the sequences discussed above. The two main routes, used to scrutinize the sentences in search of these signals, were, from a linguistic perspective, the left to right finite state grammar approach for NP identification and verb valence to cross examine the NPs and see whether or not (1) they could coalesce into one NP (21), (2) they formed two NPs but belonged to the same clause (22), or (3) they formed two NPs and belonged to different clauses (23).

(21) *student satisfaction is high.*

(22) *He gave the student satisfaction.*

(23) *If you are a student satisfaction is guaranteed.*

The clauses, once segmented, were processed and assigned to a part of speech, which could be either a noun (noun clauses) or an adverb (adverbial and relative clauses). The decision to classify relative clauses as adverbs was based on the finding that relative clauses behaved more like adverbial phrases than adjectives, as can be seen in the example below (24-26):

(24) *The man who has money.*

(25) *The man with money.*

(26) **The man rich.*

The decision when to apply the clause segmentation rules was a crucial one, because although clauses would ultimately behave as either nouns or adverbs, they could not always be treated as such in every situation. A clause, for example, can be moved in an interrogative sentence or be the subject of a verb, just like a noun, but it cannot be modified by an adjective. This restriction can be seen in the examples below (27-31), where the noun clause is sometimes allowed to fill the noun slot in a sentence and sometimes not:

(27) *Selling the company was a good idea.*

(28) *Was selling the company a good idea?*

(29) *Was the sale a good idea?*

(30) *The early sale was a good idea.*

(31) **The early selling the company was a good idea.*

This problem led to two different solutions. One was to produce some kind of restriction on the treatment of a noun derived from a clause, so that the noun would only behave as a noun when the restrictions were not present. The other solution was to delay the application of the clause segmentation rule until other rules had been applied, including, for example, the rule that coalesced nominal groups (*a very beautiful flower, the boy's gift, the brick house*) into a single noun. Although we found advantages and disadvantages in both solutions, we adopted the second one here. We felt it easier to see nominal groups as individual

units.

For the segmentation of the clauses in complex sentences, the following algorithm is proposed. Although the algorithm was constructed within an individual NLP system, it is believed that the steps are not system specific and can be applied to different approaches. The demands are that whatever system is used, it should be able to perform the following tasks, which are taken as prerequisites for the algorithm:

1. Assign part of speech to each lexical item in a sentence.

2. Look forward and backward in the sentence to solve problems like the syntactic ambiguity of the word *help* in *I need help* (noun) and *Can I help you* (verb).

3. Identify a nominal group, or, at least, find the headword in the group. The algorithm described here treats the nominal group as one segmented unit, but we believe that it can be adapted to see only the headword

The algorithm itself has three main stages. In the first stage it processes the sentence from left to right, stopping at each lexical item or nominal group — treated at this stage as a single word — and looking for explicit or implicit signals of coordination and subordination; if a signal is found, a clause initiator is attached to the lexical item at that point, and the next word is processed until the whole sentence is consumed. In the second stage the clause terminators are marked, following a similar procedure, from beginning to end of sentence. Finally, in the third stage, the clause is segmented, processed and assigned a part of speech.

The algorithm is extremely simplified but should capture all the complexity of clause segmentation. The following comments describe what happens in some of the crucial step.

Step 1: Is the item a conjunction?

The rule includes any of the explicit signals of subordination and coordination and offers no special difficulty, as long as cases of ambiguity are solved. This ambiguity involves subordinators that may belong to different parts of speech, like *as* in the following examples (32-34):

(32) *I can't run as fast as you* (adverb).

(33) *Do as required* (conjunction).

(34) *She work as a waitress* (preposition).

Coordinate conjunctions offer some additional problems: they may be ambiguous not only as regards part of speech but also in terms of what they may coordinate, that is, they may link adjectives (*strong and fast*), nouns (*mother and father*) or clauses (*He laughed and cried*). Obviously, they are clause initiators only when they link clauses, and the system has to be able to uncover them.

A further problem with coordinators is that they can link clauses with elliptical subjects, as in the examples below, where *he* is not repeated in the second clause. In this case a mechanism has to be found to recover the subject for each clause.

(35) *He laughed and cried*.

(36) *He has laughed and cried.*

(37) *He is laughing and crying.*

The solution proposed here is to retain the features of the subject (e.g. gender and number) in a buffer, to be used when necessary (In our system, in fact, these features were attached to the conjunction and used by the verb, or even adjectives, that followed it, to take decision on number, person and, sometimes, gender).

In cases where the subordinator was preceded by a preposition, both preposition and subordinator were treated as one unit, previously coalesced in our system. The following two sentences (38-39) show three examples of such co-occurrences:

(38) *This is the name / by which he was known.*

(39) *A day is the time interval from when the Sun rises to when it sets.*

This previous coalescing rule, not specified in the algorithm to save space, has some restrictions, due to the peculiarity of the English language to optionally move the preposition to the end of the clause.

One of the restrictions, although admittedly not extensively tested, was that the preposition, followed by a subordinator, could not be preceded by a verb that usually co-occurs with it. This is the case of combinations such as *turn to, refer to, rely on*, etc., which occur most often in interpolated clauses. The sentence below is an example of such constructions (40).

(40) *English is the language / that two people*

will turn to / when they cannot understand each other's tongue.

Step 2: Is the item a subjectless verb?

This rule covers both finite and non-finite verbs. The following are examples with finite verbs (41-43):

(41) *See / that the children join the program.*

(42) *If you have children / join the program.*

(43) *People / who have children / join the program.*

In (41), *children* is the subject of *join*, not the complement of any previous verb, and thus the rule does not apply. In (42), the rule applies and the word *join* initiates a new clause because it has no subject (we found it more productive to regard imperative clauses as subjectless).

In (43), the verb is preceded by an interpolated clause (*who have children*), has the word *people* as subject and obviously does not initiate a new clause. We decided, however, to apply the rule in cases like this, and provisionally mark the verb as clause initiator. We experimented with different approaches, but found this one to be the most economical, because it can be easily aborted later when the clauses are segmented.

More frequently, a verb initiates a clause when it is non-finite, as the following examples demonstrate (44-46):

(44) *I expect them / to come.*

(45) *Destroyed the house / he built another.*

(46) *Destroying the house / was easy.*

The examples look innocent and can be found in any grammar, although when authentic texts are used some problems arise. First, as pointed out above, a difference is made between verbs which constitute a verb phrase and are clause indicative, and verbs which are only part of a verb phrase, and therefore, not clause indicative. The following (47-48), for example, are not regarded as nominal verb phrases (underlined):

(47) *You should come.*

(48) *The house was destroyed.*

Second, non-finite verbs sometimes do not initiate a clause because they are preceded by a subject (49):

(49) *She wanted the boy to read the poem.*

In (49), *the boy* is not the object of *wanted* but the subject of *to read*, although *to read* is a non-finite verb; if translated into Portuguese, for example, the verb would agree in number and person with *the boy* (in fact it is the whole clause *the boy to read the poem* that is the object of *wanted* and not only the NP *the boy*).

Step 3: Is the item the subject of a verb?

This rule is applied as a last resort after all previous tests have failed. Since the purpose here is to identify the clause initiator, it is necessary to check if no other initiators are already present in the clause. In the following sentence, for example, the word *they* is the subject, but the rule is not applied because the combination from which, which precedes the subject, had already been marked as the clause initiator (50).

(50) *The books / from which they got the idea / inspired a generation.*

As seen above, when the subject is elliptical, the features of the previous subject are attached to the conjunctions. In the sentence below (51), for example, the features of *the boy* (masculine, singular, third person) are attached to the conjunction *and*, so that when the clauses are segmented, and the previous subject is out of reach, verb agreement is

solved by using the information that is attached there. The procedure is practically the same for relative pronouns when they are the subject in the relative clause (52). We suspect, in fact, that we were applying similar solutions to similar problems — in sentences that seemed to have a similar underlying structure (52-53).

(51) *The boy opened the book / and read the poem.*

(52) *The books / that arrived / were destroyed.*

(53) *The books arrived / and were destroyed.*

The immediate context surrounding the subject is obviously decisive. Notice that in the following two examples (54-55) the word *problems* can be a subject in the first sentence, thus being marked as a clause initiator, but not in the second, where it is preceded by a transitive verb.

(54) *If you work / problems show up.*

(55) *If you have problems / show up.*

Although we originally started with more than ten rules to mark clause initiation in our study, we were able to reduce them to the three above without sacrificing efficiency. When this three rules are applied, the stage is set for the second phase, which is the identification of the clause terminators. For this task, from an original set of four rules, we managed, again, to reduce the number, this time to one rule.

Step 4: Is the item followed by a clause initiator?

The rule seemed to work with all the examples we met in the *Comprehensive Grammar* and the newspaper texts, as long as one important condition is satisfied: that there is a verb phrase in the clause that is being terminated. This can be demonstrated in the example below (56). The clause terminator was put after the word *ordered*, because the following item (*arrived*) was a clause initiator.

(56) *The books / that you ordered / arrived.*

The NP *books* is also followed by a clause terminator (*that*), but the rule does not apply here because there is no verb phrase in the segment *the book*. When the clause *that you ordered* is processed, it becomes a mere

prepositional phrase attached to the NP *the books*, without any crucial syntactic function, besides that of a circumstantial adverb in our system. As the segmentation process is recursive, the whole sentence is processed again, with the relative clause now metamorphosed into a prepositional phrase. The only problem is the clause initiator, which is still there, incorrectly placed before the verb *arrived* by rule 2. The problem is only apparent, however, because, as rule 2 is recursively applied, the clause initiator disappears. The system, applying the rule incorrectly in the first moment, now recovers from the mistake, in a way, rejecting the previous hypothesis.

Step 5: Segment the clause

The process for segmenting the clause is extremely simple. It starts with the clause initiator, consumes the sentence until a verb phrase is found and then look for a clause terminator. If these three conditions are met, it is assumed that a clause was detected and the string between the initiator and the terminator is segmented. If one of the conditions is missing, it is assumed that there are no more clauses to be segmented in the sentence.

Step 6: Classify the clause

When the clause is segmented, the final task is to assign it to a part of speech. Since we have decided to consider relative clauses as adverbial phrases, there are only two possibilities here: a clause can become either a noun or an adverb, including as adverbs not only relative clauses but also all other clauses that do not pass the noun test.

We experimented with different approaches until we found that using valence information, based on subcategorizations of the verb, was the most economical. If the clause fills up the slot reserved for a noun, it was a noun; otherwise, it was an adverb. This can be demonstrated in the following examples (57-64); where we can see which clauses can be replaced by a noun and which cannot:

(57) *I know when I have time* (noun).

(58) *I work when I have time* (adverb).

(59) *He gave her what she wanted* (noun)

(60) *He gave her money.*

(61) *I know the truth.*

(62) *I work in the morning.*

(63) **I know in the morning.*

(64) **I work the truth.*

Of course a clause can also precede the verb, in which case it is even more important to classify it correctly. If it is classified as a noun it may function as the subject of the verb, entailing concord consequences, as shown in the following examples (65-66), where the criterion to decide whether the underlined clause is a noun or an adverb is based on agreement with the verb that immediately follows it. When the subordinate clause is an adverb, the verb in the main clause is in the imperative; when the subordinate clause is a noun the verb in the main clause is in third person present.

(65) *To open the doors turn the keys* (adverb)

(66) *Turning the keys opens the doors* (noun)

Notice, by the way, how concord affects grammaticality in the following examples (67-70):

(67) *What you said affected everybody.*

(68) *What you said affects everybody.*

(69) *What you said can affect everybody.*

(70) **What you said affect everybody.*

In terms of implementation, the rule used in our system to classify a clause as a noun, in this context, can be translated in the following terms:

If the clause is immediately followed by an inflected verb (third person, past tense, can, may, etc.), classify it as a noun; otherwise, classify it as

an adverb.

Testing the Algorithm

The algorithm, following the steps described here, was tested on a set of 500 sentences, randomly selected from a corpus of 10,000,000 words of expository text. The texts were about different topics and only sentences with more than one clause were selected. Table 2 summarizes how the clause types were distributed and shows the scores achieved by the system in correctly classifying the clauses.

The easiest to classify were the relative clauses, which produced a perfect score. It seems that, as regards relative clauses, the problems had all been solved by the time the algorithm was applied, including the resolution of the ambiguity involving some of the relative pronouns, mainly *that*. As can be seen in the sentences below (71), the word *that*, depending on the surrounding syntactic context, can be, respectively, a subordinate conjunction, a determiner, a relative pronoun, an adverb, and a pronoun

(71) *I didn't know <u>that</u> <u>that</u> car <u>that</u> you bought ten years ago was <u>that</u> reliable. <u>That</u> really surprises me.*

The lowest score was achieved with coordinate clauses. This was due mainly to the ambiguity of the conjunction in terms of what it was linking, as discussed above. In a sentence like (72), for example, the system had difficulty in detecting that the word *and* was linking two clauses and not just the numerals *17* and *11*. In (73), on the other hand, the system was not able to see that the underlined *and* was linking a coordinate clause, which contained a nested relative clause with another coordinate clause, containing, in turn, another nested clause. In fact, this seems to be difficult even for human beings. It took some of the students, working in the project, a lot practice to see how complex sentences were organized, and sometimes, with some sentences, they were unable to discern their underlying structure.

Clause Type	No.	Score
(In 500 sentences)		
Main clauses	402	383(95%)
Coordinate clauses	303	272(89%)

nominal clauses	226	211(93%)
relative clauses	229	229(100%)
adverbial clauses	499	484(97%)
Total	1,659	1,579(95%)

Table 2: Clause types correctly classified by the algorithm

It is not argued, however, that the algorithm was only as good as the data it started with. It does not ratify the old aphorism "garbage in, garbage out." It often improved the data it received, leading to much better results, especially with interpolated clauses and subjunctive nouns.

(72) *He began working in a steel mill at the age of 17 and 11 years later he became an organizer for the union.*

(73) *The Renaissance, which began in the 14th century, was a period of great accomplishment for European artists and architects, and the age of exploration, beginning in the 15th century, included voyages to the far corners of the world by European navigators.*

The algorithm did a better job in segmenting the clauses, splitting the sentences in the right places in more than 98% of the cases. Considering that the correct classification of a clause is very much dependent on its correct segmentation, it is not surprising that segmentation scores should be higher, although some leeway should be given for certain adverbial phrases to more or less freely move from one clause to another with very little consequences. The following sentence (74), for example, written without commas, is segmented by the system before *when*, which is probably more appropriate, but could as well be segmented after newspaper.

(74) *I read the newspaper in the evening when I have time.*

Sometimes segmentation and classification are interdependent, affecting each other. In the following example (75), shown as it was segmented by the system, the first clause (*He was irritated at Edwin*) and the last one (*he could not have said*) were both classified as independent clauses. This incorrect classification happened because the system was unable to figure out how to segment the last two clauses, which are inverted (the expected order would be: *though he could not have said where the advantage lay*).

(75) *He was irritated at Edwin / taking / what seemed to him like an unfair advantage, / though where the advantage lay / he could not have said.*

Another way to test, and indirectly evaluate the algorithm, is to compare it with other systems. This was informally done by comparing its output with the output of some machine translation programs available in the market. Examples (76-A - 77-C) show the results of this comparison, where (76-B and 77-B) were produced by our system, and (76-C, 77-C) by what is regarded, in terms of popular translation programs, as the best in the market (*Power Translator Pro 6.2*, as reviewed by Mourão, 1998). (76-C and 77-C) are not only syntactically ill-formed but also incomprehensible for native speakers of Portuguese.

(76-A) *The man who said that the boys love the girls arrived yesterday.*

(76-B) *O homem que disse que os meninos amam as meninas chegou ontem.*

(76-C) *O homem que disse que os meninos amam que as meninas chegaram ontem.*

(77-A) *His cartoons, which also appeared in the newspaper, included caricatures of presidents.*

(77-B) *Seus desenhos, que também apareceram no jornal, incluiram caricaturas de presidentes.*

(77-C) *As caricaturas dele que também apareceram no jornal caricaturas incluídas de presidentes.*

It has to be emphasized here that when the output produced by our system is compared to that of other programs, the comparison is based on selected examples involving the problems described in this investigation. It is not claimed that the program developed for this study to enter the clause segmentation rules is in any way superior to the others. In general terms, it isn't. What makes the difference is the attention given to clause segmentation.

Also, this investigation has two serious limitations, one in terms of methodology, the other in terms of the algorithm itself.

In terms of methodology, since the study was based on a rather limited lexicon, some lexical items had to be added during evaluation. Although some strict rules were followed during these additions — no proper names and new disambiguation rules, for example, were allowed to be added — it can always be argued that what was added could favor the results.

In terms of the algorithm itself, after a lot of experimentation, the processing order of the segmented clauses, not described in the steps, was first clause type (noun, relative, adverbial, coordinate, and finally the main clause) and then from left to write, when the clauses belonged to the same type. This solves many of the problems we found in the study but is admittedly primitive. The procedure was adopted due to the difficulty of grasping the underlying clause structure in complex sentences, as discussed above.

Conclusion

This investigation was conducted with the purpose of evaluating the importance of clause segmentation in NLP, exploring the property that clauses have of being encapsulated into a single part of speech and being perceived as such by the other parts of the sentence. For this purpose an algorithm was proposed and tested, using a machine translation system with examples from English and Portuguese.

We tried to make the algorithm not only as simple as possible, reducing all clauses to either an adverb or noun, but also powerful so that subtle differences in clauses and the roles they played with neighboring verbs could be adequately performed.

Bibliographical References

Borba, F. S. (1996). *Uma Gramática de Valências para o Português*. São Paulo: Ática.

Brazil, D. (1995) *A Grammar of Speech*. Oxford: University Press.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton,.

Chomsky, N. (1986). *Barriers*. Cambridge, Mass.: MIT Press.

Goodluck, H. (1991). *Language Acquisition: A Linguistic Introduction*. Oxford: Basil Blackwell.

Halliday, M. A. K. *An Introduction to Functional Grammar.* London: Edward Arnold.

Hocket, C. F. (1955). A manual of phonology. *International Journal of American Linguistics*. 21(4): Memoir no. 11.

Keenan, E. L. and Comrie, B. (1977). Noun phrase accessibility and universal grammar, *Linguistic Inquiry*, 8(1), 63--99.

Mourão, L. (1998). Fale sua língua. *PC World*. São Paulo: Feb.(68), 21--28.

Quirk, R. et al. (1985). *A Comprehensive Grammar of English.* London: Longman, 1985.

Stefanini, M. H. (1993). Talisman: Une Architecture Multi-agents pour l'Analyse duFrançais Ecrit. Thèse de Doctorat. Grenoble: Université Pierre Mendès-France, Centre de Recherche en Informatique Appliquée Appliquée aux Sciences Sociales.

Woods, W. A. (1970). Transition network grammars for natural language analysis. *Communications of the ACM*, n. 13, 591--606.