Anaphora Resolution Without World Knowledge

(Resolução da Anáfora sem Conhecimento de Mundo)

Vilson J. Leffa
(Universidade Católica de Pelotas)

Abstract: *A typical problem in the resolution of pronominal anaphora is the presence of more than one candidate for the antecedent of the pronoun. Considering two English sentences like (1) "People buy expensive cars because they offer more status" and (2) "People buy expensive cars because they want more status" we can see that the two NPs "people" and "expensive cars", from a purely syntactic perspective, are both legitimate candidates as antecedents for the pronoun "they". This problem has been traditionally solved by using world knowledge (e.g. schema theory), where, through an internal representation of the world, we "know" that cars "offer" status and people "want" status. The assumption in this paper is that the use of world knowledge does not explain how the disambiguation process works and alternative explanations should be explored. Using a knowledge poor approach (explicit information from the text rather than implicit world knowledge) the study investigates to what extent syntactic and semantic constraints can be used to resolve anaphora. For this purpose, 1,400 examples of the word "they" were randomly selected from a corpus of 10,000,000 words of expository text in English. Antecedent candidates for each case were then analyzed and classified in terms of their syntactic functions in the sentence (subject, object, etc.) and semantic features (+ human, + animate, etc.). It was found that syntactic constraints resolved 85% of the cases. When combined with semantic constraints the resolution rate rose to 98%. The implications of the findings for Natural Language Processing are discussed.*

Key-words: *Anaphora Resolution; Natural Language Processing; Textual Constraints; Ambiguity.*

Resumo: *Um problema típico na resolução da anáfora pronominal é a presença de mais de um candidato para antecedente do pronome. Considerando duas frases como (1) "As pessoas compram casas de luxo porque elas oferecem status" e (2) "As pessoas compram casas de luxo porque elas querem status", podemos perceber que os dois SNs "pessoas" e "casas de luxo", de uma perspectiva estritamente sintática, são ambos can-*

didatos legítimos para antecedente do pronome "elas". Este problema tem sido tradicionalmente resolvido pelo uso do conhecimento de mundo (Teoria de Esquemas, por exemplo), onde, através de uma representação interna do mundo, "sabemos" que casas "dão" status e que as pessoas "querem" status. O pressuposto neste trabalho é de que o uso do conhecimento de mundo não explica como o processo desambiguador funciona e explicações alternativas precisam ser exploradas. Usando uma abordagem pobre em conhecimento de mundo (informação explícita do texto em vez de conhecimento de mundo implícito) este estudo procura investigar até que ponto restrições sintáticas e semânticas podem ser usadas para resolver a anáfora. Para isso, 1.400 exemplos da palavra "they" foram aleatoriamente selecionados de um corpus de 10.000.000 de palavras de texto expositivo em língua inglesa. Os candidatos a antecedente em cada caso foram analisados e classificados de acordo com sua função sintática (sujeito, objeto, etc.) e seus traços semânticos (+ humano, + animado, etc.). Os resultados mostraram que as restrições sintáticas resolveram 85% dos casos. Quando essas restrições foram combinadas com as restrições semânticas, o percentual de resolução aumentou para 98%. Discutem-se, finalmente, as implicações desses resultados para o Processamento da Língua Natural.

Palavras-chave: *Resolução da Anáfora; Processamento da Língua Natural; Restrições Textuais; Ambigüidade.*

1. Introduction

A word can be said to have two parts: form and content. In very simple terms, this means that for every linguistic form there is at least one concept that corresponds to it. The form "tree", for example, either as a string of sounds, pronounced by somebody, or a string of letters, printed on the page, corresponds to a concept that we usually have of trees as made up of trunk, branches and leaves. The relationship between form and content ¾ signifier and signified in Saussure's terms ¾ is very close, like the two sides of a coin. Signifier and signified are unified in one larger unit, usually defined as a linguistic sign, and cannot be separated.

Obviously, when language is used by people in real-life situations, the Saussurean dichotomy, so limpid in theory, becomes fuzzy. First, there is the problem of ambiguity, where one linguistic form can refer to many different objects in the world, and vice-versa. Second, there is the much more complicated problem of anaphora, where a linguistic form does not refer directly to a concept but to another linguistic form, which may then eventually relate to a concept.

Anaphora can be described as a process that entails a go-back in the text. The process starts when the anaphor is met (e.g. a pronoun) and concludes when the antecedent is found (the word the pronoun refers to). Describing what happens between these two moments is the purpose of this paper. The goal is to offer a description to a level of explicitness that can be used for implementation in different computational languages, including Prolog, C, or Basic.

2. Tracking down the antecedent

The following segments illustrate in some detail what is involved in anaphora resolution and serve to demonstrate how the process works. Segment 1, below, chosen for its simplicity and lack of ambiguity, is used to illustrate the basic concepts underlying this process.

Segment 01: *Houses* ^{i} are bought because *they* ^{i} offer comfort.

The pronoun *they* does not relate directly to an object in the world but to a word that was mentioned before. The mental task performed by the reader, when processing this sentence, is to go back in the text and find the word it relates to. In Segment 1, there are four words preceding the pronoun (*because, bought, are* and *houses*), but only one is a serious candidate (*houses*). The pronoun *they* can only be replaced by a plural noun and the only word that fulfills this condition is *houses*.

Examples in real-life situations are not always so straightforward. One complication that may arise is the possibility of more than one rightful candidate for the antecedent, as demonstrated in the following case:

Segment 02: *Houses* ^{i} are bought by people because *they* offer ^{i} comfort.

Now there is not one but two candidates for *they*, which are the words *houses* and *people* (both plural nouns). How to solve this problem? One hypothesis is that it can be solved by applying syntactical constraints. It can be argued that there is a syntactic parallelism between the noun *houses* and the pronoun *they*, that is, both *houses* and *they* are in the subject position in their own clauses. The word *people*, on the other hand, although a plural noun, does not share this parallelism with *they*. Thus, between the two candidates, we choose the noun *houses*.

Syntactic constraints based on parallelism, however, seem to work fine only as far as the examples are carefully chosen. In Segment 2, for example, simply changing one lexical item for another may totally revert the relationship between the anaphor and the referent. This can be seen in the following Segment, where *offer* is replaced by *like.*

Segment 03: *Houses* ^{i} are bought by *people* ^{j} because *they* ^{j} like comfort.

Again there are two candidates, which are exactly the same as in the previous Segments. But if we apply syntactic constraints, as we did before, choosing the noun phrase that is in the subject position, we would end up with the word *houses*, which obviously is the wrong choice (**Houses like comfort*). Syntactic parallelism, which so efficiently facilitated the choice between the two candidates in the previous Segment, does not seem to work any longer. The only candidate that can rightfully be in the position occupied by *they* is *people*. The other candidate (*houses*) violates a semantic constraint: *houses* do not *like* things; only *people* like things. In other words, the verb *like* requires for subject a noun with the semantic feature +ANIMATE. Syntactic parallelism can, thus, be overruled by semantic constraints; it is not enough for the antecedent to possess the same syntactic function as the anaphor; both antecedent and anaphor must also share semantic features.

Syntactic and semantic constraints, still, are not enough in solving the problems associated with anaphora resolution, as can be seen in the following cases.

Segment 04: The *companies* ^{i} sold their cars ^{j} to the sheiks ^{k} because *they* ^{i} offered long-term guarantee.

Segment 05: The companies ^{i} sold their *cars* ^{j} to the sheiks ^{k} because *they* ^{j} were bulletproof models.

Segment 06: The companies ^{i} sold their cars ^{j} to the *sheiks* ^{k} because *they* ^{k} offered more money.

Segments 4-6, apparently, can only be solved by resorting to a representation of the world in which there are sellers, buyers, and commodities that change hands: money from buyers to sellers and cars from sellers to buyers. We also need to know that cars can be bulletproof, that companies usually offer guarantees on what they sell and that sheiks can be very rich. All this world knowledge has to be readily available for the antecedent of *they* to be correctly identified in each of the Segments.

The problem, however, in using world knowledge is its computational cost. There are so many variables that a combinatory explosion is inevitable. Each variable can interact with so many other variables, with so many possibilities of different combinations that the system may enter an endless loop and the right combination is never encountered.

The solution to the problem of tracking down the antecedent in anaphora seems to lie somewhere between the simplicity of syntactic constraints and the complexity of world knowledge. This is the problem addressed in this investigation. There are two questions to be answered here: (1) what are the limitations of syntactic constraints in anaphora resolution? (2) what other possible solutions can be found between these constraints and world knowledge?

3. Discourse, cognition and textual constraints

Anaphora can be studied from different perspectives, including discourse (e.g. McEnery and Botley, 1998; Indursky, 1997), cognition (e.g. Langacker, 1996; van Hoek, 1992) and textual constraints (Dagani and Itai, 1990; Nasukawa, 1994; Mitkov and Belguith, 1998). Many of these studies emphasize the correlation between certain discourse-pragmatic factors (e.g. topicality) and a given anaphoric form (reference-tracking device in the terminology of Du Bois, 1980). Fox (1996) summarizes these correlations as follows:

(a) use of pronouns or zero when anaphors are closer to the topic being developed; use of full nouns when topicality is low;

(b) use of pronouns or zero when anaphors are in the same discourse sequence of previous mentions; use of full nouns when they are not;

(c) use of pronouns or zero when speaker assumes hearer is paying close attention; use of full nouns when speaker assumes low level of attention;

(d) use of pronouns or zero when speaker is not emotionally involved;

(e) use of full nouns when speaker's attitude is highly positive or negative. (Fox, 1996:vii)

Reference-tracking devices such as the use of pronouns, zero anaphors or full nouns, even when correlated to topicality, discourse sequences and speaker's cognitive or affective states, do not reveal very much in terms of the process involved. All it amounts to, in fact, is the probability of which anaphor to use - on a scale that ranges from zero to a full noun. There is no description of what really happens in the mind when the reader or listener finds an anaphor and tries to track down its antecedent, inside or outside the text. Anaphora resolution at this low level of processing, most of it below conscious control, is probably not an area of interest to the discourse-pragmatic paradigms of research, which may be more concerned with the general picture, viewing the process from a more abstract level.

A very different perspective is offered by studies in computational linguistics, where the implementation of an anaphora-resolution system makes it necessary to translate abstract concepts into machine-readable code, using data that have to be found on the textual surface. With the processing power of modern computers, the variety of these data have been increased. We are no longer limited to low-level types of linguistic data, such as part of speech information, but we can also include higher-level linguistic structures, related to possible configurations of relationships between different segments of the text. We can recursively encapsulate chunks of language into ever-increasing units, building larger blocks, and abstracting their characteristics. The crucial point, however, is that the link between the anaphor and the antecedent should be unambiguous, with total agreement between different readers consuming the same text. Should disagreement arise, not due to differences in the texts but to differences in readers' interpretations, the problem is beyond solution from the perspective of computational linguistics, which is basically algorithmic. Attempts to endow computers with the world knowledge needed to attribute meanings to text, instead of only extracting meaning from it, are theoretically interesting but extremely costly and, for the time being, unfeasible for practical purposes. Anaphora resolution, in terms of computational linguistics, cannot be attributed to the cognitive or affective states of the readers; the data have to be present on the surface of the text.

Linguistic data that can be found on the text such as gender and number agreement, c-command restrictions, syntactic parallelism, lexical repetition, or antecedent proximity are favored in the resolution process because they can be more easily dealt with by the available tools in computational linguistics. These tools usually rely on the concepts of "constraints" and "preferences" - where "constraint" is the more powerful of the two devices. Knowledge-poor solutions, using corpus-driven methodologies, statistical and probabilistic models, are substantially preferred.

Some strategical approaches for tracking the antecedent, as opposed to purely statistical models, have also been proposed. These approaches can be formalized in terms of rules, usually based on constraints and preferences. The following preferences, for example, can be used to select the antecedent (based on Mitkov (1994,1996)):

- The NP is the object of one of the following verbs: discuss, present, illustrate, summarize, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyze, synthesize, study, survey, deal, cover;
- The NP is modified by one of the following verbal adjectives: defined, called, so-called;
- The NP is modified by one the following adverbs: particularly, especially, namely, specially;
- The NP is the object of one the following nouns: chapter, section, table, figure, paper, report;
- The NP is repeated throughout the discourse section;
- The NP occurs in the heading of the section.

Paraboni (1997) also adopts a strategical approach, using a combination of constraints and preferences in his study of Portuguese anaphora involving possessive adjectives. These adjectives, when belonging to the third person, are interesting in Portuguese because as anaphors they do not agree in number and gender with the antecedent - as they do in English, for example - but with the thing being possessed, a feature that makes it more difficult to locate the antecedent. Reference-tracking strategies, therefore, have fewer constraints and preferences to rely on. In Paraboni´s study, for example, very few rules could be found to track down the antecedent. One of the most productive was the presence of a coordinative conjunction between the anaphor and the NP, as shown in Segment 7.

Segment 7: The law ^{i} and its ^{i} consequences

Paraboni, himself, however, is cautious to point out, that exceptions to this rule can be easily found. This seems to be the case, for example, with complex NPs such as (8) and (9), where the coordinative conjunction rule is overridden by semantic constraints (See also Baltazart and Kister, 1996).

Segment 8: The book ^{i} on divorce ^{j} and its ^{j} consequences

Segment 9: The book ^{i} on divorce ^{j} and its ^{i} author.

In summary, it seems that a description of what happens in anaphora resolution falls into two extremes, offering two alternatives. On one hand, the process is analyzed in very general terms, from a highly abstract level, providing the whole picture, but failing to offer a description of crucial aspects of the process, which are implicitly acknowledged, but not explicitly detailed. On the other extreme, the process can be minutely described in one of its aspects, analyzing every little step, but then failing to provide the whole picture.

This investigation opts for a third alternative, relying heavily on the data available from the surface of the text, but, looking at them both horizontally (from a syntagmatic perspective) and vertically (in a more paradigmatic way), where semantic relations between the lexical items that make up the text are also taken into account. In Segments 8 and 9 above, for example, it can be argued that the lexical item *author* shares more semantic features with *book* than *divorce* - which should offer a cue for solving the problem of finding the right antecedent for the possessive *its*. The main purpose here is to test a procedure that can be summarized in two steps: the first is to select an antecedent that complies with syntactic constraints (including gender, number, syntactic parallelism, etc.); the second, and more complex, step is to check for semantic constraints.

4. Methods

Anaphora resolution is a crucial issue in Natural Language Processing (NLP). Any project in the area of computational linguistics, including information retrieval, dialogue processing, and machine translation, has to allocate a major part of the system to solve this problem. The decision at what stage of the process to attack the problem is dependent on many aspects, including the theoretical approach being used. For the approach proposed here, based on a machine translation project between English and Portuguese, anaphora is dealt with after some preliminary analysis has already been performed on the text being processed, including the following:

Part of speech assignment: each word in the text has already been classified into one of the basic word classes (noun, verb, adjective, etc.) and subclasses (transitive verb, intransitive verb, etc.).

Attachment of specific attributes: number (singular, plural), semantic features (+HUMAN, +ANIMATE, etc.), and gender specifications that were necessary for the Portuguese translation (masculine, feminine) were also attached to the NP.

Noun phrase segmentation: complex nouns such as combinations of two nouns (*stone house*), adjective and nouns (*the big house*) have already been segmented with identification of the corresponding headword. The segmentation also includes combinations of more than one NP such as *the president of the United States, George Bush, and England's Prime Minister, Tony Blair*, which forms a complex plural NP.

Case assignment: the syntactic function (nominative, accusative, dative, etc.) of the resulting NP is already known.

Table 1 shows how two NPs are classified. Notice that *a large house on the hills* is classified as singular NP, since all the words that make up the noun phrase are governed by the headword *house.*

Table 1 - Noun phrase segmentation

Tourists	**prefer**	**a large house on the hills**
NP		Noun
Masculine		Feminine
Plural		Singular
(+) Animate		(-) Animate
Nominative		Accusative

For this investigation the pronoun "they" was chosen. There is a theoretical and a practical reason for this choice. In terms of theory, it is expected that the analysis will help explain the interrelation between anaphora and text, from a strictly linguistic point of view. The question asked here is whether or not it is possible to resolve anaphora without resorting to world knowledge - basically how far anaphora is dependent on syntactic and semantic constraints. In terms of practice, the results could be immediately applied to machine translation from the English language into many romance languages such as French, Spanish or Portuguese, where the pronoun *they* has different translations, depending on the antecedent.

The basic methodology involved a selection of 1,400 occurrences of *they* from a corpus of 10 million words of explanatory texts. For this selection a concordancer program was used. This type of program allows for the occurrence of a given word or combination of words to be automatically extracted from the corpus and listed according to a selected order (alphabetically by first left word, first right word, second word, etc.), thus facilitating different analyses.

After the 1,400 segments were selected, the antecedent was identified and classified in accordance with its syntactic function (subject, direct object, indirect object, etc.). In Segment 10, for instance, the antecedent is *the Aztecs* and has the function of subject.

Segment 10: Continually dislodged by the **small city-states**[h] that fought one another in **shifting alliances**[i], **the Aztecs**[j] finally found refuge on a small island in Lake Texcoco where, about 1345, ***they***[j] founded the town of Tenochtitlan.

The semantic features of the verb that followed the pronoun were also analyzed, in terms of the traits they required in the subject. This can be shown in Segment 11, where there are 7 candidates for the antecedent of *they* (*economists, solutions, problems, economies, markets, prices* and *exports*), but only the NP *economists* can be chosen because, although the furthest from the anaphor, it is the only one that can be the subject of *cite* without producing a semantic anomaly.

Segment 11: **Economists**[g] who disagree with imposed **solutions**[h] to Third World development **problems**[i] point to the excessive vulnerability of Southern **economies**[j], which are largely dependent for their growth upon relatively open Northern **markets**[k] and reasonable international **prices**[l] for their **exports**[m]. ***They***[g] cite the need to involve local populations (...).

The practical methodology used to uncover semantic constraints was simply to line up all the possible candidates from the closest to the furthest, starting from the anaphor, until an acceptable antecedent was found. This is shown below - taken from Segment (11) - where the adequate NP is found only at the 7th attempt. The verb *cite* can only accept as subject a noun, which has the +HUMAN semantic feature.

They cite the need to involve local populations.

1. * **exports** cite the need to involve local populations.
2. * **prices** cite the need to involve local populations.
3. * **markets** cite the need to involve local populations.
4. * **economies** cite the need to involve local populations.
5. * **problems** cite the need to involve local populations.
6. * **solutions** cite the need to involve local populations.
7. **Economists** cite the need to involve local populations.

An algorithm-like heuristics was used to detect the syntactic and semantic constraints available in the text, as summarized below:

Step 1 Look for a plural NP to the left of *they*, up to 80 words in the text. If an NP is found, go to Step 2. If not (in the 80-word stretch of text), go to Step 4.

Step 2 Does the NP have the same syntactic function as *they*? If the answer is *yes*, go to step 3; if *no*, go back to step 1.

Step 3 Can the NP replace *they* without producing semantic anomaly? If the answer is *yes*, go to step 7; if *no*, go back to step 1.

Step 4 Look for a plural NP to the left of *they*, up to 80 words in the text. This step is only taken if the 80-word limit is found without meeting the condition in Step 2 (syntactic function). The procedure starts again, this time considering only semantic constraints. Thus, if an NP is found, go to Step 5. If not (in the 80-word stretch of text), go to Step 6.

Step 5 Can the NP replace *they* without producing semantic anomaly? If the answer is *yes*, go to step 7; if *no*, go back to step 4.

Step 6 No solution found. If no NP is found in the 80-word limit, adopt a default procedure (e.g. Translate *they* as masculine). Go to Step 7.

Step 7 Finish procedure. Look for further occurrences of anaphors in the following segments.

The procedure is divided into two testing phases, each of them leading to a solution if the candidate for antecedent passes the syntactic and semantic tests. Using Segment 11 to demonstrate the syntactic phase, we can see that all the candidates in the passage, with the exception of *economists,* do not pass Step 2, which means that they are discarded at the syntactic level (They lack syntactic parallelism for not sharing the same subject function with the anaphor). Only the NP *economists* reaches Step 3. As it passes the test, steps 4, 5 and 6 are ignored and, in this case, the anaphora is resolved.

It should be noted that in the procedure proposed here syntactic parallelism by itself (subject/subject) is not qualified to decide whether or not a candidate NP can be classified as the antecedent for an anaphor; this decision can only be taken at the semantic level. Syntactic parallelism is therefore subjected to semantic constraints. Step 3 is the first of these decision points: if a solution is found, the procedure is finished; if not, the procedure resumes, going back to Step 1. The process is repeated until the 80th word to the left or a full noun in the subject position is encountered.

The semantic testing phase is only activated if no NP passes Step 3. This can be demonstrated through Segment 12. Since no solution was found considering both syntactic and semantic constraints, a second round starts now, ignoring syntactic constraints.

Segment 12: An amnesty is an exemption from prosecution for criminal acts, usually issued by a government after a time of crisis such as a war or revolution. The amnesty may be for acts such as rebellion, treason, desertion, or draft evasion. It is usually granted to groups of **citizens**^{j} on condition that ***they***^{j} abide by the law in the future.

The first round ends, in this case, when the beginning of the passage is found. The second round starts and hits on *citizens* as the first plural NP.

It is not a subject, but since syntactic constraints are no longer taken into account, the NP is only tested for semantic anomalies and passes the test.

In case the NP is not a noun but a plural pronoun, the procedure goes on, looking for a full noun, until the 80-word limit or a singural subject is found. This can be seen in Segment 13, where the process, starting from the last *they*, passes the pronoun *they* (in *They tried*), and stops at (*Mongol bands*).

Segment 13: Following Kublai Khan's eventual overthrow of China's Song dynasty in 1279, **Mongol bands**^{i} raided much of Eastern Asia outside of China. ***They***^{j} tried in vain to invade Japan in 1274 and 1281, captured Burma's Pagan in 1287, and penetrated Champa and Annam in 1285-88. ***They***^{j} even attempted to invade Java in 1292-93.

When the procedure described above is unable to find an antecedent for the anaphor, it is marked as unresolved, and a default value may be used. This can be seen in Segment 14, for example. The procedure would be unable to find a plural NP - which in this case happens to be a combination of a subject (*Perseus*) with an object (*Andromeda*).

Segment 14: When Cassiopeia boasted that Andromeda was more beautiful than the sea-goddess called Nereid, Poseidon, god of the sea and Nereid´s father, sent a sea monster to ravage Ethiopia. Only the sacrifice of Andromeda could persuade Poseidon to call off the monster, so Andromeda was chained naked to a sea cliff. The hero Perseus saw her plight, rescued her, and killed the monster. Thereupon, Poseidon turned the dead monster into the sea's first coral. **Perseus**^{j} married **Andromeda**^{j}, and ***they***^{j} eventually became king and queen of the Greek city of Tiryns.

5. Results and discussion

This investigation attempts to answer three questions: (1) how many cases were solved by applying only syntactic parallelism, considering only cases that do not affect semantic constraints, that is, stopping at the first phase of the procedure? (2) how many cases were solved by applying only semantic constraints, therefore, ignoring syntactic parallelism; and finally (3) how many cases were not solved. Table 2 shows the results, in terms of percentage.

Table 2 - Success rates of two different tracking devices

Tracking device	%
Syntactic Parallelism	86
Semantic Constraints	12
Unresolved	2

Syntactic parallelism is the strongest factor, solving alone, 86% of the cases. This means that simply looking for a plural subject, ignoring semantic constraints, leaves only 14% of the cases unresolved.

If semantic constraints are taken into account, however, 12% more of the cases are resolved, raising the percentage to 98%. A review of the literature reporting on investigations that used syntactic and semantic constraints combined with statistical approaches shows that this is the highest percentage ever obtained. Table 3 summarizes the results attained by some of these studies in pronominal anaphor resolution, not only in English but also in Polish and Arabic.

Table 3 - Success rates in anaphor resolution

Study	%
Baldwin (1997)	75
Mitkov (1998) (English)	89.7
Mitkov (1998) (Polish)	93.3
Mitkov (1996)	94.7
Mitkov and Belguith (1998)	95.2
Mitkov (1998) (Arabic)	95.2
Mitkov and Stys (1997)	95.8

This percentage obtained in our study should be surprising, especially if we consider that the procedure used here is far simpler than the ones used in other studies, sometimes combining complex scales of preferences and statistical approaches along with syntactic and semantic constraints.

A possible explanation is that the pronoun “they” may be easy in terms of antecedent tracking when compared to other pronouns. For one thing, *they*, in the vast majority of cases, refers to NPs explicitly mentioned in the previous segment of text; unlike *it*, for example, which may be an expletive, refer to a whole sentence or a previous paragraph, instead of simply referring to an NP. Also, the antecedent of an anaphor tends to be the focus of the paragraph, and, since the focus tends to be in the subject position, that would, again, facilitate finding the antecedent for *they*.

Syntactic parallelism (subject co-referring to subject) was found to be very powerful, sometimes overriding paragraph focus. It was felt that the antecedent of "they" could be found in any type of clause including main clauses, subordinate clauses and even *interpolated* clauses. This can be demonstrated in Segment 15, where the NP *founders*, although in a subordinate clause, is the rightful antecedent for *they* and would be correctly selected by the algorithm, simply because it is the first plural NP to occupy the subject position.

Segment 15: **Historians**^{i} continue to debate what the nation's **founders**^{j} meant to include when ***they***^{j} wrote that there shall be "no law" abridging the freedom of speech or press,

I would like to argue, however, that the high rate of resolution is due to a successful combination and ordering of syntactic preferences and semantic constraints as used in the proposed algorithm. In fact, if the semantic constraints were not applied at the exact moment the NP in the subject position is found, the results would be very different.

By considering only syntactic parallelism, 94% (not 86%) of the segments investigated would satisfy the condition, but that would produce an error margin of 14% (instead of the 2%). This can be demonstrated by analyzing Segment 16: by applying only syntactic parallelism, the selected antecedent would be the NP *farmers*, because, like *they*, the NP is in the subject position. The choice of *farmers*, however, would be incorrect because the right antecedent is *chickens*, although it lacks syntactic parallelism, for being in the direct object position. Semantic constraints, based on the verb *purchase*, would favor *chickens* more than *farmers* - as merchandise are more likely to be purchased than people - and so, in accordance with the proposed algorithm, chickens would be correctly chosen.

Segment 16: In specially constructed **broiler sheds**^{i}, the **farmers**^{j} raise the **chickens**^{K} to market weight, at which point ***they***^{k} are purchased.

Admittedly, I may have stretched the application of semantic constraints here a little beyond the limits of what is usually done in NLP systems. I assume it is feasible to provide lexical items with conditional semantic traits to be assigned at the moment of processing by applying some collocational rules, as demonstrated in Segment 17; both the object in active sentences (*we purchase things*) and the subject in passive sentences (*things are purchased*) should have the semantic trait -HUMAN. I think this is within the limits of most NLP systems. I may be expecting too much, however, when I combine semantic traits of the verb with semantic traits of both object and subject, as in Segment 17, where I assume the system would be able to choose correctly between "quasars lie mainly toward the edge of the known universe" and "*astronomers lie mainly toward the edge of the known universe".

> Segment 17: Because the objects called **quasars**^{i} exhibit **large red shifts**^{j}, **most astronomers**^{K} think that ***they***^{i} lie mainly toward the edge of the known universe.

5.1. Unresolved cases

Applying syntactic rules, combined with semantic constraints, leads to a resolution rate of 98%, which seems to indicate a successful procedure. It would be interesting, however, to examine the remaining 2% that were not resolved and see if an explanation can be found for them. It seems that the unresolved cases can be grouped under two conditions: (1) The antecedent was incorrectly chosen by the algorithm and (2) the antecedent could not be found. In Segment 18, we have an example where the algorithm, using syntactic parallelism, and finding no semantic constraint, made an incorrect choice by selecting *subsistence economies*, which is in the subject position, instead of *couple*, which, in spite of being in the object position, is the rightful antecedent. The algorithm, although able to detect, based on labels assigned to the lexical items, that *couple* was a plural noun (like *people*, *government*, *family*, etc.), was not sophisticated enough to detect the correct clues that led to *subsistence economies.*

> Segment 18: **Subsistence economies**^{i} could provide any given **couple**^{j} with access to **goods**^{k} and **services**^{l} that ***they***^{j} alone could not provide.

Sometimes the antecedent was not found simply because there was no plural NP in the previous segment. This seemed to be the case where some kind of summation was involved, as demonstrated in Segment 19. The algorithm was unable to detect Perseus, the subject of the sentence, and Andromeda, the object of the verb, as making up a plural NP.

Segment 19: Perseus married Andromeda, and ***they*** eventually became king and queen of the Greek city of Tiryns.

The challenge to create a more abstract and economical rule that could encapsulate these 2% of unresolved cases, without including world knowledge, led to many rewritings of the algorithm and countless returns to the literature on anaphora, but proved to be fruitless, and I had to leave it as it is. It could be argued, as consolation, that the inability to solve 2% of the cases does not look bad if we consider that human beings, using all their knowledge of the world and years of experience with the language, do not seem to fare much better. I found, working with university students who acted as research assistants in this project, that they themselves were often unable to locate the right antecedent. What does look bad, however, is the system's inability to solve cases that are extremely easy for any speaker of the language to solve, as is the case in Segment 19. Further attempts to rewrite the algorithm to solve these cases created more problems than it solved, and, once again, I had to stop there. Anaphora resolution with 100% accuracy remains, in my view, a Holy Grail in Natural Language Processing.

6. Conclusion

Anaphora resolution, using only syntactic and semantic constraints, without resorting to encyclopedic or world knowledge, has both a bright and a dark side. The bright side is the high rate of success, which may reach percentages above 95%, close to fluent speakers of the language. Quantitatively, the results, even if not impressive, can be interpreted as very good. The dark side is on the quality of the mistakes produced, which are sometimes ridiculous from a world knowledge perspective based on common sense and human intuition.

The temptation is to claim that there is much more in anaphora resolution than can be seen from the data collected on the surface of the text; world knowledge seems to be the only reliable source, after all. Resorting to world knowledge, however, in terms of Natural Language Processing by computers, in my view, is just transferring the problem to a higher level of abstraction without solving it. Common sense, intuition, socio-historical variables, and other components of world knowledge are all too evasive and vague to be adequately treated in terms of Computational Linguistics.

A solution to avoid ridiculous mistakes has to lie beyond syntactic constraints based on gender and number agreement or other syntactic parallelisms between anaphors and antecedents, such as subjects with subjects, direct objects with direct objects and so on - but cannot go as far as what has been vaguely defined as world knowledge; the constraints are untreatable at that level. Possible paths that should be explored here include the concept of collocation - starting with Firth's idea that a word is known by the company it keeps, and including the contribution of Hoey (1991) on patterns of lexical repetition, where the emphasis is more on lexical than on grammatical relations. Charolles's (1988) metarules, exploring the need for combinatorial order and logical connections between the lexical items in the text, could also be useful.

Any solution brought to anaphora can contribute to other areas of language study such as ambiguity resolution, textual cohesion and, eventually on reading comprehension and text production. The relationship between anaphora and ambiguity, for example, is so close that it is probably impossible to refer to one without using the other; anaphora itself is a type of ambiguity. This is also true of textual cohesion, considering that a discourse is a logical sequence of ideas tied together according to certain preferences and constraints. In more practical terms, we can also argue that the findings of studies on anaphora will eventually contribute to reading and writing instruction, by showing the students the mechanisms used to connect different parts of the text.

E-mail: leffa@via-rs.net

Recebido em dezembro de 2001
Aprovado em fevereiro de 2003

REFERENCES

BALDWIN, B. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. Trabalho apresentado no *ACL'97/ EACL'97 workshop on Operational factors in practical, robust anaphora resolution*. Madrid, 11 July, 1997.

BALTAZART, D. and L. KISTER 1996. Is it possible to predetermine a referent included in a French N de N structure? In S. P. BOTLEY and A. M. MCENNERY (eds.) *Discourse anaphora and anaphor resolution colloquium*. Lancaster: Lancaster University.

CHAROLLES, M. 1988. Introdução aos problemas da coerência dos textos (abordagem teórica e estudo das práticas pedagógicas). In C. GALVES, E. ORLANDI and P. OTONI. (eds) *O texto: escrita e leitur*a. Campinas: Pontes.

DAGAN, I. & ITAI, A. 1990. Automatic processing of large corpora for the resolution of anaphora references. *Proceedings of the 13th International Conference on Computational Linguistics*, COLING'90, Helsinki.

DU BOIS, J. 1980. Beyond definiteness: the trace of identity in discourse. In W. CHAFE (ed.) *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.: 203-74.

FOX, B. Introduction. 1996. In B. FOX (ed.) *Studies in anaphora*. Amsterdam: John Benjamins: vii-xi.

HOEK, K van. 1992. *Paths through conceptual structure: Constraints on pronominal anaphora.* Doctoral dissertation. San Diego: University of California.

HOEY, M. 1991. *Patterns of lexis in tex*t. Oxford: University Press.

INDURSKY, F. 1997. Da anáfora textual à anáfora discursiva. *Anais do 1º. Encontro do Círculo de Estudos Lingüísticos do Sul – CelSul* vol.2. Florianópolis: UFSC: 713-22.

LANGACKER, R. W. 1996. Conceptual groupings and pronominal anaphora. In B. FOX (ed.) *Studies in anaphora*. Amsterdam: John Benjamins: 333-78.

MCENERY, T. and S. BOTLEY (eds.) 1998. *Discourse anaphora and anaphor resolution*. Amsterdam: John Benjamins.

MITKOV R. 1996. Anaphora resolution: a combination of linguistic and statistical approaches. *Proceedings of the Discourse Anaphora and Anaphor Resolution*. Lancaster University, UK:17-19.

MITKOV, R. 1998. Robust pronoun resolution with limited knowledge. *Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*. Montreal: Canada.

MITKOV, R. 1994. A new approach for tracking center. In *Proceedings of the International Conference New Methods in Language Processing*, UMIST, Manchester, UK:13-16 September.

MITKOV, R. and L. BELGUITH 1998. Pronoun resolution made simple: a robust, knowledge-poor approach in action. *Proceedings of the International Conference Traduction Automatique et Langage Naturel (TALN'98*). Paris: France.

MITKOV, R. & M. STYS 1997. Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'97)*: 74-81. Tzigov Chark, Bulgaria.

NASUKAWA, T. 1994. Robust method of pronoun resolution using full-text information. *Proceedings of the 15th International Conference on Computational Linguistics* COLING'94, Kyoto, Japan: 5-9 August.

PARABONI, I. 1997. *Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa*. Dissertação de mestrado. Porto Alegre: PUCRS.