

Using Local Focus to Correct Illegal NP Omissions¹
(A Ph.D. Proposal)

Linda Z. Suri

Technical Report No. 93-07
March, 1992

Linda Z. Suri
Department of Computer and Information Sciences
103 Smith Hall
University of Delaware
Newark, DE 19716
suri@udel.edu

¹This research was supported in part by NSF Grant #IRI-9010112. Support was also provided by the Nemours Foundation.

1 Introduction

Correcting text which is ill-formed with respect to grammar and/or discourse strategies is a challenging problem. We are working on this problem from the perspective of helping deaf writers produce text which conforms to the standard rules of English.² This perspective may prove to be particularly interesting since the native language of some deaf writers is American Sign Language (ASL), which differs from English in both its syntax and its discourse strategies and thus may have an interesting influence on written English.

ASL is a visual-gestural language whose grammar is distinct and independent of the grammar of English or any other spoken language [Sto60], [Lid80], [Ing78], [BP78], [BC80], [Bak80], [Pad81], [Pad82], [HS83], [KB79], [BPB83]. In addition to sign order rules, ASL syntax includes systematic modulations to signs as well as non-manual behavior for morphological and grammatical purposes [BC80], [Lid80], [Pad81], [KB79], [KG83], [Ing78], [Bak80]. The modality of ASL encourages simultaneous communication of information which is not possible with the completely sequential nature of written English.

The work described in this proposal is part of a much larger project. The long term goal of the overall project is to develop an instructional writing tool which will take a writing sample from a deaf person, analyze it to identify deviations from standard English, engage the user in a corrective tutorial dialogue, and generate text which is correct with respect to the context. The overall system can be seen as having two phases. The *identification phase* will rely on a grammar of English which has been augmented with a set of syntactic and semantic error productions ([Sle82], [WS83], [WVJ78]) which extend the language accepted by the parser and semantic interpreter to include the types of deviations we expect. The *interactive tutorial and correction phase* will be driven by annotations on the error productions as well as discourse information which will be tracked through the dialogue.

The work being proposed in this document is part of the correction and tutorial phase. In particular, it focuses on discourse information which must be tracked in order to generate a correction to a particular class of errors. The particular solution that we are proposing is motivated by the hypothesis that the underlying source of these errors is the transfer of a discourse strategy from ASL to English. This hypothesis is substantiated by an analysis of writing samples and a comparison of ASL and English which has led us to conclude that language transfer, if defined broadly enough, can explain many of the errors we have uncovered³. This explanation has led us to an algorithm for correcting the particular class of errors we are concentrating on in this thesis. The algorithm works for every instance we have so far uncovered in our samples.

The proposed thesis work focuses on one error class which we have found to be particularly prevalent and interesting: the illegal omission of (both pleonastic⁴ and contentful) NP's. The question under investigation for the proposed thesis is how these omissions can be corrected. In this proposal, we first substantiate the claim of language transfer by briefly describing the sample analysis which motivated our current beliefs. This includes a brief exploration of LT (see [Sur91] for a much more thorough background on LT), a characterization of how LT might manifest itself, and a description of the results of the analysis.

²This tool would be very useful to the deaf population. Since data on writing skills is not well-documented, we note that the reading comprehension level of deaf students is considerably lower than that of their hearing counterparts, "... with about half of the population of deaf 18-year-olds reading at or below a fourth grade level and only about 10% reading above the eighth grade level..."[Str88]

³It should be noted that we have not attempted to *prove* that language transfer is behind the errors we have found. Rather, we will show that language transfer is a *reasonable explanation*.

⁴A pleonastic NP is one which does not play a thematic role. For example, *it* in "It is raining" or "I like it when Mary sings", and *there* in "There is a book on the chair".

While previous work on LT has mostly concentrated on the sentence level, we hypothesize that LT may also occur at the discourse level. That is, not only may a writer transfer syntactic structures and lexical items from a native to a second language, but discourse and cohesion strategies may be transferred as well (and we believe our analysis substantiates this claim).

We describe classes of ASL verbs and a discourse strategy of ASL. The rich verb agreement of one class of verbs explains some empty categories (EC's) in ASL, and the discourse strategy (Topic-NP Deletion) explains EC's occurring with ASL verbs with no agreement morphology. We use these facts about ASL (and the existence of other instances of LT between ASL and English) to explain illegal NP omissions in deaf writing samples.

Next, we explain how we can use this information about ASL to correct illegal NP omissions in deaf written English. Basically, the analysis of ASL verbs and EC's indicates that the topic is most likely to be the referent of the omitted NP. Thus an algorithm that tracks topic might successfully fill in the missing NP's. We provide examples of illegal NP omissions in deaf writing samples which we think can be explained by this analysis. Having chosen this approach to illegal NP correction, we discuss focus (topic) tracking research which has been done in NLP. We explain why we chose one of these (Sidner's focus tracking algorithm ([Sid79], [Sid83])) as the basis for the algorithm we will use.

Next, we describe the proposed focusing algorithm, and show how the algorithm predicts the correct missing referent for three deaf writing samples with illegal NP omissions. Each of the examples motivates extensions that were needed by Sidner's algorithm. We discuss these along with some open questions about how to further extend the algorithm.

The major contribution of this thesis will be the development of the focusing algorithm. The tracking of focus is important to Natural Language Processing since the recognition of focus is important for understanding discourse, and it is particularly important for anaphora (including ellipsis) resolution. Thus, this focusing algorithm should be useful not only for the larger project of correcting the written English of ASL natives, but for other NLP tasks as well.

2 Writing Sample Analysis

The purpose of this section is to describe the writing sample analysis which has motivated our conjecture that many errors in the writing of ASL natives can be explained as LT. We first describe some background on LT and then explain our analysis.

2.1 LT

The term Language Transfer is commonly used to refer to the use of knowledge of one's native language (L1) in the production and/or comprehension of a second language (L2). Transfer may be positive (in the sense that it may speed the acquisition of the L2), however it may also result in deviations in L2 production in places where the L1 and L2 differ. While the existence of LT has been a rather controversial subject over the years (see [McL87],[GS83a],[Sur91]), much recent research has provided convincing evidence of LT resulting in the transfer of L1 lexical items, syntax rules and pragmatic production rules to L2 (e.g., [Sch82] and [SR79] as described in [McL87]; [Kle77], [Hak76], and [Gas79] as described in [Gas84]; [GS83a]; and [McL87]).

Given that transfer has been documented between spoken languages, one might ask whether or not LT could occur between ASL (a visual-gestural language) and written English.⁵ At first glance, transfer may seem surprising since the components of ASL grammar and written English grammar

⁵Other researchers (e.g., [PQ73], [QSW74], [QWM76], [RQP76], [QPS77], [KK78] [QP84]) studied errors in deaf writing but did not attribute errors to LT.

are different [Sto60], [BP78], [Pad82], [HS83], [KB79], [BPB83]. ASL grammar components include sign order, morphological modulations of signs, and non-manual behavior which occurs simultaneously with the manual signs [BC80], [Lid80], [Pad81], [KG83], [Ing78], [Bak80]. Written English grammar components include word order, morphological modulations of words, and punctuation, but nothing that clearly corresponds to the simultaneous non-manual behavior found in ASL. On the surface, the fact that ASL and written English occur in different modalities seems problematic as well. However, research shows that much of ASL processing occurs on the side of the brain primarily used for processing spoken and written languages (the left-side of the brain), as opposed to the (right) side of the brain primarily used for visual and spatial functions [Sac90]. Thus, we expect transfer is likely to occur between ASL and written English, but it is not immediately clear how the transfer will manifest itself (particularly with respect to the non-manual component of ASL grammar).

2.1.1 A Characterization of LT

Because of the differences (in grammar and modality) between ASL and English, we have attempted to abstractly characterize how languages could differ in a way which is independent of the grammar components. We have identified several ways in which languages may differ which might lead to (negative) transfer.

- Two languages may differ in *when* they mark a particular feature. As a result the marking of that feature in the L2 may seem either redundant or overly concise/imprecise in the native language. For example, in ASL it is usual to establish tense at the beginning of a discourse, and then not to mark it again until the time frame changes. Of course, in English tense is marked (on the verb) in every finite clause. So, marking tense in every finite clause in English may seem redundant to the ASL native. Transfer of such a feature (i.e., when to mark tense) might explain omission errors (in this case, of tense markings) in the written English of ASL natives.
- Two languages may differ in *how* they mark a feature. For example, in ASL, Yes/No questions are distinguished from declarative statements with non-manual markers (facial expression and body shifts). This is radically different from the word order changes which typically mark Yes/No questions in written English. Thus LT might explain errors in Yes/No question formation by the ASL user.
- Languages differ in regard to requiring morphological changes or additional lexical items for strictly syntactic reasons. For example, adding an “s” to a present tense verb for a third person singular subject in English (typically) conveys no extra information, but is a syntactic requirement. There is not a close counterpart to this in ASL, which may explain omissions of this morphological item in the written English of ASL natives.
- As with any two languages, English often has two or more words or phrases which correspond to a single ASL sign (or sign sequence), and vice versa. For example, ASL uses the same sign (i.e., lexical item) for “other” and “another”. Thus, LT might explain why ASL learners of written English might take some time to learn which word (“other” or “another”) to use in English.

2.1.2 Examples of Error Classes Attributable to LT

We collected writing samples from a number of different schools and organizations for the deaf. We concentrated on eliciting samples from Deaf people who are (native) ASL signers. This was done

in order to increase the probability of finding errors specific to the deaf population and to allow us to target a more homogeneous population.

Thus far, we have analyzed forty-eight Freshman and Sophomore writing evaluation samples from Gallaudet University, a liberal arts university for the deaf, seventeen writing evaluation samples from the National Technical Institute for the Deaf (NTID), twelve first draft papers from students in the high school program at the Margaret S. Sterck School, a deaf school in Delaware, and five letters and essays written by ASL natives contacted through the Bicultural Center in Washington, DC. The total sample size is approximately 25,000 words.

<ul style="list-style-type: none"> ● Conjunctions: 4 <ul style="list-style-type: none"> – Omitted conjunction: 1 – Inappropriate conjunction: 3 ● Prepositions: 66 <ul style="list-style-type: none"> – Omitted preposition: 26 – Inappropriate preposition: 27 – Extra Preposition: 13 ● Determiners: 63 <ul style="list-style-type: none"> – Omitted determiner: 35 – Inappropriate determiner: 9 – Extra Determiner: 19 ● Incorrect Number on Noun: 23 ● Incorrect Subject-Verb Agreement: 11 ● Tense and Aspect: 70 <ul style="list-style-type: none"> – Dropped tense: 5 – Other tense/aspect problems: 65 ● Mixing up English words or phrases which share a single ASL sign: 12 ● BE, HAVE (non-Auxiliary): 16 <ul style="list-style-type: none"> – Omitted BE: 9 – Lack of BE/HAVE distinction: 7 ● Other Omitted Main Verbs: 7 ● Incorrect WH-phrase: 4 ● Adjective Problems: 13 <ul style="list-style-type: none"> – Incorrect Adjective Choice: 3 – Incorrect Adjective Formation: 10 ● Incorrect Nominalization: 5 	<ul style="list-style-type: none"> ● Relative Clauses: 14 <ul style="list-style-type: none"> – Relative pronoun deletion: 4 – Resumptive pronoun: 1 – Other: 9 ● Pronouns: 12 <ul style="list-style-type: none"> – Incorrect pronoun choice (including pleonastic): 7 – Inappropriate pronoun use (where full definite descriptions are required): 4 – Lack of pronoun use (overuse of definite descriptions): 1 ● Pleonastic Pronoun Deletion: 10 <ul style="list-style-type: none"> – Object: 5 – Subject: 5 ● Focus/Discourse Structuring Problems: 49 <ul style="list-style-type: none"> – Omission of focused element (subject: 4; object: 4): 8 – Problems carrying over general/specific description strategies: 5 – Structuring problems with “because”: 8 – Ambiguous modifier attachment: 1 – Other (possibly related to carry-over of topic-comment strategies): 27 ● Redundancy Problems: 2 ● Not Enough Sentence Breaks: 6 ● Other: 104 (23% of errors in database)
--	---

Table 1: Error Taxonomy

Table 1 contains the error taxonomy⁶ we have derived from the analysis of 79 writing samples. Also included in the table is the number of sentences (out of 214 sentences) which contained a deviation which could be explained by each classification. These numbers are based on only the 17 samples (3313 words) which have already been added into our database.⁷ The intent of the

⁶Error classes which occur less frequently have been classified under “Other”.

⁷We have recently created a database to store analyzed writing samples. A database user can retrieve all sentences

numbers is to give the reader an idea of how often the various classes occur in relation to each other. In Appendix B, we give an example for each error class listed in Table 1.

We will show how many of the error classes uncovered in our analysis can be explained as following from one (or more) of the 4 above categories of differences between ASL and written English, and thus may be explained by LT. For each illustrated error class, we provide examples of the error class and then explain how it could be captured by our characterization above. (More detail can be found in [Sur91].)

Conjunctions

- Conjunction Deletion:
 - “He taught _ directed, for almost 30 years ...”⁸

While researchers have identified several kinds of conjunctive markings in ASL from body shifts to particular lexical items ([Pad81], [BS88]), there are many places where an explicit conjunction would be required in written English, but not in ASL. For instance, conjoined verbs do not require an explicit separate lexical item; instead the verbs are signed one after the other [Fan83]. Therefore, it is not surprising that an ASL signer would omit ‘and’ between (the final and next-to-final) conjoined verbs in written English. This omission could be the result of the marking seeming redundant or radically different to the ASL native.

Subject-Verb Agreement

- Incorrect Subject-Verb Agreement:
 - “*My brother like to go...*”

In ASL, not all verbs mark subject agreement for person and number. For certain verbs (some directional and classifier verbs) subject agreement is indicated by a change in handshape, a change in movement, or (rarely) the use of an overt NP where it would not normally be needed [Pad81]. There is a large class of verbs in ASL which do *not* vary in form for person and number of the subject (see [Pad81]). In addition, some directional verbs vary in form according to the person and number of the *object* [BC80], [Fan83].

That subject-verb agreement is a syntactic constraint in English, coupled with the difference in *when* and *how* agreement is marked in the two languages, might explain deviations in marking subject-verb agreement in the written English of ASL natives.

Tense and Aspect

- Dropping tense on verbs (within or across sentences):
 - “We *went* to see Senator Biden’s office ... Then we *go to see* the Vietnam memorial”
- General tense/aspect problems:

(with their corresponding corrected sentences) containing a particular error. Entering the data into the database is very time-consuming, which is why only 17 samples have been entered thus far.

⁸**Note:** “_” is used to mark places where we think the writer has omitted one or more words from the corresponding correct English sentence.

- “Many students rather *live* at college, than *living* at home.” (Correction: “*Many students would rather live at college than live at home.*”)

In our data we found missing and incorrect tense markings, and missing and incorrect aspect markings. These deviations might be explained by the differences between when and how ASL and English mark tense and aspect.

Some ways that English marks tense and aspect are through the use of modals and auxiliary verbs, and through morphological changes to the verbal elements. ASL does not use auxiliaries and it does not modulate verb signs for tense.⁹ In ASL, tense is generally established once at the beginning of a discourse (e.g., by using a time sign), and that time frame is understood to persist until the next time frame is established.

The dropping of verb tenses in the written English of the deaf might be explained by this; marking tense by modulating every verb in English might seem redundant to someone fluent in ASL. Similarly, the problems we found with the formation of English verb tenses might also be explained by the fact that ASL marks time in a radically different manner from the way that English does (i.e., using a specific time indicator instead of adding phonemes and auxiliary verbs).

English also uses the sequential addition of auxiliaries, prepositional phrases, adverbs, etc. for aspectual distinctions. ASL signs are modified, *through changes in movement of the underlying sign*, for aspectual changes [KB79]. Adverbial modification is often achieved through facial expression [BS88]. Since the methods of achieving aspectual distinctions in English are so radically different from those of ASL, the above aspectual deviations can be explained by LT.

BE, HAVE

- missing BE:
 - “Once the situation changes they _ different people.”
- lack of distinction between “be” and “have” (as main verbs):
 - “... some birth controls *are* side-effect.” (Correction: “... *have side-effects...*”)
 - “I wish to go to Hawaii because it *is* beautiful *and nice weather.*” (Ellipsis implies “... and *is* nice weather.” Correction: “... and *has* nice weather”)

ASL does not have a “be” sign which could explain its omission in the written language of the deaf. In ASL, the idea of *being* is conveyed by radically different means, for instance, by a topic-comment structure. Generally, a topic is set up, and then properties are attributed to the topic (the topic and comment are distinguished non-manually).

While ASL does have a “have” sign, it is often omitted if it can be assumed from the context.

Lexical Items

- Mixing up English words or phrases which share a single ASL sign:
 - “Somehow, *I am interesting in ASL* and I want to learn it.”

A single sign in ASL corresponds to both “interesting” and “interested.” Thus this error is attributable to LT.

⁹The positioning of a verb with respect to the ASL time line may reflect tense.

2.2 LT Summary

While we have only explained a few examples, at least 82% of our error codes (which represent a finer distinction than that given in Table 1), and at least 76% of the individual errors reported in Table 1, are attributable to LT in a similar manner.¹⁰ That so many errors fit the characterization above suggests LT is an important predictor of errors and we should use this characterization to hypothesize new error classes. Our characterization can be useful in developing language tutoring systems for second language learners of other languages as well.

2.3 Discourse Level Errors and LT

We believe that several of the error classes we have uncovered in the written English of ASL natives are the result of transfer of *discourse strategies* from ASL to written English. We term the resulting errors *discourse errors*. They manifest themselves either at the sentence level (resulting in a sentence which is ill-formed syntactically or semantically), or they may only be apparent in a longer stretch of text.

Wilbur [Wil77] reports that the language instruction of the deaf has concentrated on the sentence level and suggests that many errors could result from the writer not understanding when and how to use particular structures (such as relative clauses and pronouns). We believe that the concentration on sentence level teaching may contribute to the discourse errors we have found, and it may result in discourse level errors persisting longer than sentence level errors. Discussions with educators of the deaf and ASL researchers confirmed the idea that a skilled deaf writer may develop his/her writing skills to a point where he or she produces text which lacks discourse cohesion, even though the individual sentences are grammatically correct.

Much of the research on LT has concentrated on how differences in L1 and L2 syntax affect the surface syntactic form of L2 production. Other researchers have explored the effects of features which are more discourse-related; we summarize their work here.

Odlin claims that comprehension and production problems in an L2 may arise due to lack of familiarity with a discourse pattern, or lack of familiarity with culturally specific knowledge. See [Odl89] for many examples of such problems. Some particularly interesting examples are the less accurate recollection by Americans (as compared with Japanese) of information in a passage written in an indirect form (common in Japanese) [Odl89]; differences in value judgments of writing due to cultural differences with respect to indirectness [Odl89]; and comprehension problems for British students caused by the inclusion of supernatural events in a story from the Kathlamet [Odl89]. These examples point to the facts that different cultures organize information differently, different languages exhibit significantly different levels of directness, and culturally specific knowledge may play a greater role in comprehension than one might expect.

Zobl believed LT could occur as the prolongation of *the use of L1 pragmatic strategies* involving given versus new information [GS83a]. Thus, LT might result in overuse of a particular pragmatic strategy in an L2 that does not make much use of that particular pragmatic strategy. Koch and Bartelt have documented similar transfer with respect to overuse of repetition. Koch saw evidence suggesting that Arabic discourse may encourage Arab students to repeat words and phrases in English [Odl89]; and Bartelt saw evidence of overuse of repetition in the English writing of Navajo and Apache students. [Odl89]

Rutherford argues that whether one's native language is topic-prominent, subject-prominent, or neither, and the extent to which one's native language tends to use word order to express

¹⁰We do not claim that each instance of an error class that is attributable to LT necessarily resulted from LT, only that LT could explain the error, and thus *may* be the source of the error. We recognize that there are other sources for errors, including incorrect analyses of English on the part of the writer, and English instruction.

pragmatic information, influences L2 production. The first feature seems to influence how often one uses topic-comment structures, and the second influences use of dummy subjects (in languages that have dummy subjects). We might also see lack of use of pragmatic strategies or lack of sensitivity to pragmatic strategies in the L2 production of a language which is heavily influenced by pragmatics by someone whose native language is less influenced by pragmatics.

Gass ([GS83b], [Gas89]) argued that one must learn not only the possible word orders of an L2, but to what extent surface word order is determined by pragmatic factors and to what extent it is ordered by grammatical factors. She studied the L2 production of Italian speakers learning English, and English speakers learning Italian. In Italian, surface word order is largely determined by pragmatics; in English, surface word order is primarily determined by grammatical relationships. She noted that only Italian speakers at advanced levels of English acquisition seemed to recognize the importance of syntax in English.

The work described above suggests that LT can explain the effects of an L1 on an L2 in terms of word order, use or lack of use of dummy-subjects, use or lack of use of topic-comment structures, use or lack of use of pragmatic strategies, indirectness, repetitiveness, and use of cultural information.

Thus, we propose that differences in language structuring and cohesion strategies must also be examined as sources of potential difficulty for referent formation and anaphora resolution. We will briefly describe the differences between ASL and English with respect to some discourse strategies, in order to explain how we believe LT may manifest itself between ASL and written English in terms of referent formation and anaphora resolution.

3 ASL and English Discourse Strategies and Deletions

3.1 Loci and inflecting verbs

Before describing ASL discourse strategies, it is important to understand two related aspects of ASL - establishing a locus in space as associated with an NP or a referent, and inflecting verbs for agreement. (The following description is largely based on [LM91] since she so clearly and concisely described these features of ASL.)

In ASL, the *locus* associated with a referent that is present is the location of the referent. Pronominal reference to a present referent is achieved by indicating its locus. A locus may be indicated by pointing or gazing at the locus, or by using that locus as the starting or end point of a verb. Reference to 1st, 2nd and (a present) 3rd person are achieved in this manner.

For referents which are not present, abstract loci can be associated for each referent in the signing space in front of the signer's body. "This is accomplished by producing the sign for the referent at some arbitrary locus in space, or making the sign and then pointing to the locus with the index finger, or by eyegaze in the direction of the locus while making the sign." (p. 25-6, [LM91]) Subsequent pronominal reference to a non-present referent assigned to a locus is achieved in the same manner as that for pronominal reference to present referents. Abstract loci associations persist in discourse until a new framework is established, and the number of such loci, while theoretically unlimited, typically does not exceed 5.

ASL has four types of verbs: 1) inflecting, inflectable or agreeing verbs, 2) plain or non-agreeing verbs, 3) spatial, and 4) classifier verbs [LM91]. (We will discuss only the first two.) Inflecting verbs have very rich subject and object agreement morphology, and plain verbs have no agreement morphology for subject and object.

Inflecting verbs are those that are marked for subject and/or object agreement. Subject or object agreement is achieved by a change in the movement of the underlying sign. Specifically, the movement of the sign begins at the locus of the subject, and ends at the locus of the object.¹¹ For example, if one signs, “John kicks Mary”, the starting point of the sign KICK is at the locus for John, and the ending point of the sign is at the locus for Mary. “Thus, ASL verbs do not indicate the common person and number agreement, but agreement with actual referents.” (p. 29, [LM91])

Padden claims that subject agreement is optional, though there may be restrictions on the optionality [LM91]. If the verb is not inflected for the subject, then the starting point of the verb is in a neutral location in front of the signer’s body. On the other hand, object agreement is obligatory if the locus for the object has been established.

Plain or non-agreeing verbs do not change in any way based on the subject or object.

3.2 Null Argument Structures in ASL

Lillo-Martin [LM91] identifies two kinds of null-argument structures in ASL: those involving inflecting verbs, and those involving plain verbs.

The inflecting verbs in ASL are rich in morphology, like verbs in Italian (which is considered a pro-drop language). Lillo-Martin argues that when an empty category occurs with such a verb (that is, when a subject or object is not explicitly signed for such a verb), the referent of the empty category (EC) is determined by the agreement morphology. In terms of ASL, the agreement morphology causes the verb sign to begin at the locus of the subject, and end at the locus of the object. The arguments of the verb can thus be recovered from these locations. Thus, Lillo-Martin’s analysis of EC’s with respect to inflecting verbs in ASL is consistent with the analysis of EC’s in Pro-drop languages (which share the characteristic of having strong morphological markings).

The plain verbs of ASL have no morphology, and thus one would not expect dropped NP’s because the referents couldn’t be recovered through verb morphology. However, these verbs do allow null arguments in some contexts. The same is true of Chinese verbs. Lillo-Martin argues that the deletions with respect to these verbs are similar to deletions in other languages (like Chinese) termed *discourse-oriented* languages which do not contain morphological markings yet do allow NP deletions that can be recovered from context. Her analysis of deletions/EC’s with respect to plain verbs in ASL is based on Huang’s analysis of EC’s in Chinese. Specifically, Lillo-Martin argues that ASL allows *Topic NP Deletion*, i.e., for the topic of a sentence to be “deleted under identity with a topic of a preceding sentence” [Hua84], and that ASL plain verbs may have an EC subject/object as the result of topic movement. Thus, as Huang argues for Chinese verbs, null argument structures of plain verbs arise when the signer topicalizes the sentence, thus creating an EC subject/object coindexed with the topic, and deletes the topic under coindexation with a topic of a preceding sentence.

Languages that allow Topic NP Deletion are said to be *discourse-oriented languages*. Discourse-oriented languages are very sensitive to pragmatics. English is said to be a *sentence-oriented language* [LM91]. Sentence-oriented languages do not allow Topic NP Deletion.

The point of central importance is that the deletion of an NP that co-refers with the topic of a previous sentence (or discourse topic) is not permissible in English, though it often is in ASL.

¹¹There are verbs for which the movement begins at the direct object and ends at the object. Padden [Pad88] refers to these as *Backwards verbs*.

4 Transfer of Discourse strategies from ASL to English

The central hypothesis of the proposed thesis is that the differences between ASL and English at the discourse level may explain some of the cohesion errors in deaf writing noted both in our initial analysis and informally by others. Of particular interest to us are the NP deletion errors which might result from the writer carrying the discourse-oriented aspects of ASL over to English even though English is a sentence-oriented language. In this case, the writer appears to believe that the NP can be omitted because it is a topic of the preceding text, even though English does not allow such omissions.

Examples of NP omissions which we think are related to the transfer of ASL discourse strategies include the following¹²:

- “I think that Gallaudet College should require all deaf students to take speech and speechreading courses. Therefore, they can improve their oral skills for their future use. I am going to tell you that why the deaf student should take _”
- “There are many things I like about NTID. They offer supporting services like interpreters and notetakers for mainstream classes which I had experiences through my public schools. Now NTID/RIT offers same thing that my school offered but only better supporting services. That is why I like about NTID.
But one thing worries me that most about NTID/RIT is financial problems. I hope I could find some ways to solve _.”
- “First, in summer I live at home with my parents. I can budget money easily. I did not spend lot of money at home because at home we have lot of good foods, I ate lot of foods. While living at college I spend lot of money because _ go out to eat almost everyday. At home, sometimes my parents gave me some money right away when I need _. While in college, I could not ask my parents for money right away because I live in Washington DC and my parents live in Illinois. It is too far.”

For each of the above examples, discussions with ASL informants confirm that the omitted items would be understood, and that the corresponding ASL discourse would be acceptable/grammatical if the omitted NP were not signed, pronominally referenced, or indicated by verb agreement. We propose that these (and other NP deletions like them) can be corrected if we track the topic, or, in computational linguistics terms, *the local focus* of discourse. We propose to do this by developing a modified version of Sidner’s focus tracking algorithm [Sid79], [Sid83]. Sidner’s algorithm tracks both a local focus and an *actor focus*. A claim of this work is that the deleted referent can be recovered by using focusing data structures and rules similar to those developed by Sidner for recovering referents of definite pronouns and definite noun phrases.

5 Focusing Algorithm

5.1 Related Work on Focusing

There have been several other approaches to tracking focus of one kind or another through a discourse. Grosz [Gro77], Grosz and Sidner [GS86], and McCoy and Cheng [MC88] describe algorithms for tracking *discourse* focus, as opposed to local focus. Discourse focus is intended to capture the

¹²These are actual excerpts. Each typically contains several errors. Here we focus on the deleted NP’s only.

CF	Current Focus
AF	Actor Focus
PFL	Potential Focus List
PAFL	Potential Actor Focus List
EC	Empty Category

Figure 1: Key to Abbreviations

broad set of things that the discourse is about. These algorithms are concerned with a level of focusing which is too global for tracking the omitted NP’s. This belief is supported by tests of our local focusing algorithm on sample texts. In addition, Grosz’s work relies on a task model and requires that the structure of the discourse reflect the structure of the task. The discourses that we must handle do not fall under “task oriented” dialogues for which the task model is evident. While we cannot use either Grosz and Sidner’s or McCoy and Cheng’s models in our work, we should note that Sidner’s algorithm and our extensions to her algorithm are intended to track local focus, and are consistent with these other models.

[?] describe an approach for tracking local focus. While the model is similar to Sidner’s model in that it has a *backward looking center* which roughly corresponds to Sidner’s CF, and a set of *forward looking centers* which roughly corresponds to Sidner’s potential focus list, the model (as described) leaves many unanswered questions. For instance, it does not include a record of past forward and backward looking centers corresponding to the focus stack in Sidner’s algorithm. The stacking mechanism (which we have found the need to extend beyond what Sidner specified) is necessary for recovering the referents of deleted NP’s.

[Dah86] describes a focusing algorithm which uses syntactic rather than thematic criteria to determine focus preferences. Sidner’s algorithm uses thematic, syntactic and pronominal information to determine focus preferences, and we have found these focus preferences, in the context of our expansion of Sidner’s algorithm, to be useful in correcting illegal NP omissions. Thus, we have chosen to explore expansion of Sidner’s algorithm (which is relatively well-defined) rather than exploring expansion of another approach. This approach has shown success thus far.

We should note that while Sidner’s algorithm requires inferencing to confirm the co-specification of an anaphor, any system will need inferencing at least to confirm a co-specification, and confirming a co-specification by inferencing is far easier than selecting a co-specifier by inferencing (as other systems have done). Since Sidner’s approach relies heavily on linguistic knowledge, it also relieves us of the burden of representing and reasoning with a significant amount of world knowledge beyond that needed for confirming the co-specification of an anaphor.

5.2 Overview of our Algorithm

Our focusing algorithm is basically what is described in [Sid79] and [Sid83] to track local and actor foci, but we have had to augment the algorithm to handle complex sentence types, and to track some additional information.

The proposed focusing algorithm works by recording certain information as a discourse progresses from one sentence to the next. In each (simple) sentence, the actor focus (AF) is identified with the thematic agent of the sentence. (If the sentence has no agent, then the previous AF is retained.) The Potential Actor Focus List (PAFL) contains all NP’s that specify an animate element of the database and do not occur in agent position. If a sentence has a pronoun in agent

position, the previous AF is chosen for its co-specification.¹³

Tracking local focus requires some additional machinery. The first sentence in a text can be said to be *about* something. That something is generally different from the actor focus and is called the current focus (CF) of the sentence.¹⁴ The CF can generally be identified via syntactic means, taking into consideration the thematic roles of the elements in the sentence (see Appendix A for description of algorithm). In addition to the CF, an initial sentence introduces a number of other items (any of which can go on to become the focus of the next sentence). Thus, these items are recorded in a potential focus list (PFL).¹⁵

After the first sentence, at any given point in a well-formed text, the writer has a number of options:

- Continue talking about the same thing; in this case, the CF doesn't change.
- Talk about something just introduced; in this case, the CF becomes a member of the previous sentence's PFL.
- Return to a topic of previous discussion. In this case that topic must have been the CF of a previous sentence, or must have been on the PFL of a previous sentence.¹⁶

The decision (by the reader/hearer/algorithm) as to which of these alternatives was taken is based on the thematic roles (with particular attention to the agent role) held by the anaphoric elements of the current sentence, and whether their co-specification is the CF of the previous sentence, a member of the PFL of the previous sentence, or an element in the CF stack or the PFL stack. Confirmation of co-specifications requires inferencing based on general knowledge and semantics.

At each sentence in the discourse, the CF and PFL of the previous sentence are stacked for the possibility of subsequent return. When one of these items is returned to, the stacked CF's and PFL's above it are popped, and are thus no longer available for return.

5.3 Filling in a Missing NP

We propose using information from the focus algorithm to identify the referent of an illegally omitted NP (and extending the focus algorithm to calculate the CF and PFL in the presence of an illegally omitted NP).

To identify the referent of a missing NP, we treat the omitted NP (whose position in the sentence will be identified by syntactic analysis) as an anaphor which, like Sidner's treatment of full definite NP's and personal pronouns, co-specifies with an element recorded by the focusing algorithm. We define preferences among the focus data structures which are similar to Sidner's [Sid79], [Sid83]. Essentially, we expect the omitted item to co-specify with a focus item of higher priority than the co-specifier of a pronoun or definite NP.

More specifically, when we encounter an omitted NP in other than agent position, we first try to fill the deleted NP with the CF of the immediately preceding sentence. If semantics and general

¹³Throughout this paper, we discuss whether an NP *co-specifies* another NP, the Current Focus, or a member of a Potential Focus List, etc. By writing that X co-specifies Y, we mean that the knowledge network representation specified by X is the knowledge network representation specified by Y (in the case that Y is an NP), or corresponding to Y (in the case that Y is a focusing data structure).

¹⁴Sometimes the AF and the CF are the same.

¹⁵Sidner uses many terms and data structures to describe her algorithms. We will collapse these terms for simplicity. For instance, Sidner writes of a PFL, ALFL, and DEF, and we will refer to all of them as a PFL. She refers to an expected focus and a current focus (CF), but we will call them both CF (even though the the CF of a discourse-initial sentence may never be confirmed as the focus, but only be expected to be the focus when processing the next sentence).

¹⁶Sidner's algorithm just stacked CF's. We have extended the algorithm to stack PFL's as well.

knowledge inferencing cause this co-specification to be rejected, we then consider members of the PFL of the previous sentence for filling the deleted NP. If these too are rejected, we consider stacked CF's and elements of stacked PFL's, taking into account preferences (yet to be determined) among these elements. When we find an element that is acceptable according to syntax, semantics and general knowledge, we fill the empty NP with that element.

When we encounter an omitted NP in agent position in a simple sentence or a sentence-initial clause, we first test the previous the AF as co-specifier, then members of the PAFL, the previous CF, and finally stacked AF's, CF's and PAFL's. To identify the missing agent NP in a non-sentence-initial clause, our algorithm will first test the agent of the previous clause, and then follow the same preferences just given. Further preferences are yet to be determined, including those between the stacked AF, stacked PAFL, and stacked CF.

While filling in an NP is much like finding the co-specifier of any other anaphor, we place the additional constraint that a missing NP in a clause should be filled before the co-specifiers of other anaphora are calculated. We impose this constraint because of the following. First, we are assuming that an omitted NP is the most focused element in the sentence. Second, we are assuming¹⁷ that if there is an omitted NP in a clause, all NP's that co-reference that NP in that clause are also omitted (since we think they could also be omitted for the same reason that the first omitted NP was omitted, i.e., they will be understood from the context).¹⁸ Thus, if we were to first compute the co-specifier of another anaphor, that anaphor would be assigned the most focused possible co-specifier, and the omitted NP would not be able to co-specify with that most focused co-specifier (under the second assumption) and will thus be forced to co-specify with a less co-specified item (violating the first assumption).

5.4 Computing the CF

We must decide how to compute the CF in the presence of illegally omitted NP's. As is specified in the algorithm mentioned in section 5.2, the CF of a sentence (in a coherent discourse) will be related to the elements contained in the data structures maintained by the focusing algorithm: it will be the same as the CF of the last sentence, an item introduced in the last sentence and thus a member of the PFL of the last sentence, or an element of the stacked CF's or stacked PFL's. The decisions as to which one of these moves is taken is determined by the anaphoric elements in the sentence and their co-specifications. When computing the CF, we treat illegally omitted NP's as anaphora since they (implicitly) co-specify something in the preceding discourse. Sidner's algorithm orders the focusing data structures, giving preference to the previous CF and then the previous PFL, and finally considering the focus stacks, and takes the first such element that has a co-specifier in the current sentence. Exceptions to this ordering occur for either thematic reasons or due to the type of anaphor used. In keeping the AF and CF different, the algorithm prefers a non-agent anaphor co-specifying a PFL member over an agent co-specifying the CF. The anaphora themselves are prioritized: pronouns are considered better indicators of focus than definite NP's. Thus, the algorithm prefers a PFL member co-specified by a pronoun to the CF co-specified by a full definite description.

In determining how the algorithm should compute the CF in the presence of omitted NP's, it is important to remember that discourse-oriented languages allow deletions of NP's *that are the topic of the discourse*. Thus there is strong evidence that a deleted NP (in the writing of an ASL native)

¹⁷We need to confirm that this assumption is reasonable based on further study of ASL and analysis of deaf writing samples.

¹⁸Recall we are only correcting NP's that are omitted under Topic NP deletion, as opposed to those deleted in the presence of rich verb morphology.

is the intended topic.

Note that Sidner prefers pronouns to a definite descriptions as the likely CF since pronouns are strong indicators of focus. We include that preference, and add the preference that we prefer omitted NP to pronouns since they are yet stronger indicators of focus (at least in discourse-oriented languages).

Thus, in adapting Sidner’s algorithm to handle the omitted NP’s, we want to prefer the deleted non-agent as the focus, as long as it closely ties to the previous sentence. Thus, we prefer the co-specifier of the omitted non-agent NP as the (new) CF if it co-specifies with either the last CF or a member of the last PFL. If the omitted NP is in agent position, we prefer for the new CF to be a pronominal (or, as a second choice, full definite description) non-agent anaphor co-specifying either the last CF or a member of the last PFL (allowing the deleted agent NP to be the AF and keeping the AF and CF different).¹⁹ If no anaphor meets these criteria, then the members of the CF and PFL focus stacks will be considered, testing a co-specifier of the omitted NP before co-specifiers of pronouns and definite descriptions at each stack level.

The description above applies to simple sentences. When we have a complex sentence, we compute the CF and PFL for each clause as if the clause occurred as a simple sentence in isolation, and then use this information from each clause to compute the CF and PFL of the entire sentence (as briefly described in section 8). This aspect of the algorithm will be one of the major contributions of this thesis.

A fuller description of the Focusing Algorithm can be found in Appendix A.

6 Overview of the Algorithms

The algorithm that we will implement and use to track focus and fill in missing NP’s is composed of several smaller algorithms (two of which were discussed in sections 5.4 and 5.3). A high-level description of the modules of this algorithm is in figure 2.

The Discourse Initial Algorithm selects the CF and PFL for the first sentence of a discourse. Next a loop is entered to process the remaining sentences in the discourse. The Co-Specification Algorithm takes the CF, AF, PFL, PAFL and focus stacks and uses a set of preferences to determine the co-specifiers of the definite NP’s, definite pronouns, and empty NP’s in the current sentence (discussed in section 5.3). We should note that the implemented algorithm will only compute the co-specifiers of the empty anaphora. We will not calculate the non-empty anaphora, but we will assume we are given their co-specifications via a set of oracles. Such co-specifications could be calculated by anaphora resolution algorithms similar to Sidner’s definite anaphora resolution algorithms. We will not reproduce Sidner’s anaphora resolution algorithms here; basically they impose preferences among the focusing data structures and the kinds of co-specification relationships that an anaphor could have with each of the data structures. (Carter extended Sidner’s algorithms to handle intrasentential anaphora.)

The Focus Tracking algorithm (partially discussed in section 5.4) calculates the new CF for each (non-discourse-initial) sentence, and stacks the previous CF and PFL. The Actor Focus Algorithm selects and AF and PAFL, and stacks the previous AF and PAFL. The Potential Focus Algorithm calculates the PFL for non-discourse-initial sentences. More complete descriptions of these algorithms can be found in Appendix A. The remaining thesis work will continue to flesh out these algorithms. For instance, the algorithms in the appendix do not handle the complex sentence and related extensions that will be discussed below.

¹⁹As future work, we will explore how to resolve more than one non-agent anaphor in a sentence co-specifying PFL elements.

Discourse Initial Algorithm - % establishes CF, PFL, AF, PAFL of the first sentence
 LOOP
 Co-Specification Algorithm - % selects co-specifications using the CF, AF, PFL,
 % PAFL, and focus stacks from the previous sentence
 Focus Tracking Algorithm - % updates the CF and CF stack to reflect the current sentence
 Actor Focus Algorithm - % updates the AF, PAFL, AF stack and PAFL stack
 % to reflect the current sentence
 PFL Calculation - % updates the PFL and PFL stack to reflect the current sentence
 GOTO LOOP

Figure 2: Flow of Algorithm Processing

7 Example 1

Below, we describe the behavior of the extended algorithm on an example from our collected texts containing both a deleted object and deleted subject.

Example:

(S1) First, in summer I live at home with my parents.
(S2) I can budget money easily.
(S3) I did not spend lot of money at home because at home we have lot of good foods, I ate lot of foods.
(S4) While living at college I spend lot of money because __ go out to eat almost everyday.
(S5) At home, sometimes my parents gave me some money right away when I need __.
(S6) While in college, I could not ask my parents for money right away because I live in Washington DC and my parents live in Illinois. ”

After the Discourse Initial Algorithm is applied to S1, the CF is **HOME**, and the PFL is **SUMMER**, and the **LIVE VP**, the AF is **I**, and the stacks are empty.

Focus Data Structures after S1:

CF	HOME
PFL	SUMMER, and the LIVE VP
AF	I
CF stack	empty
PFL stack	empty

For S2, we first apply the co-specification algorithm. Next, the Focus Tracking Algorithm is applied. I is the only anaphor, so it becomes the CF, the PFL is **MONEY**, **EASILY**, and the **BUDGET VP**, the AF is **I**, the CF stack contains **HOME**, and the PFL stack contains the previous PFL.

Focus Data Structures after S2:

CF	I
PFL	MONEY, EASILY, and the BUDGET VP
AF	I
CF stack	HOME
PFL stack	SUMMER, and the LIVE VP

S3 is a complex sentence using the conjunction “because.” Such sentences are not explicitly handled by Sidner’s algorithm. As noted earlier, we will compute the CF and PFL for each clause of a complex sentence as if it were a simple sentence following the preceding sentence, and then calculate the CF and PFL of the whole sentence based on those calculations.²⁰ For “X BECAUSE Y” sentences, we prefer elements of the X clause as focus candidates to those of the Y clause (see section 8). Thus, we take the CF from the main clause, and rank elements in the main clause before elements in the second clause on the PFL.²¹ In this case, the co-specification algorithm will identify several anaphora: “I”, “money”, and “at home”. The CF becomes **MONEY** since it co-specifies with a member of the PFL and since the co-specifier with the last CF (I) is the agent of S3. Because the PFL algorithm will order the elements of the main clause before the elements in the other clause (the one after “because”), the PFL will contain **HOME**, **NOT SPEND VP**, and **GOOD FOOD**, **HOME**, and the **HAVE VP**. The AF remains I. We stack the CF, AF and the PFL of S2.

Focus Data Structures after S3:

CF	MONEY
PFL	HOME , NOT SPEND VP , and GOOD FOOD , HOME , and the HAVE VP .
AF	I
CF stack	I , HOME
PFL stack	PFL of S2, followed by the PFL stack of S2

Note that S4 has a missing agent in the second clause. To identify the missing agent in a non-sentence-initial clause, our co-specification algorithm (which fills empty NP’s) will first test the AF agent of the last clause for possible co-specification. Since this poses no contradiction, the omitted NP is filled with “I”. The CF is computed by first considering the first clause of S4, since the X clause is the preferred clause of an X BECAUSE Y construct. Since “money” co-specifies with the CF of S3, and nothing else in the preferred clause co-specifies a member of the PFL, **MONEY** remains the CF. The PFL contains **COLLEGE**, the **SPEND VP**, **ALMOST EVERY DAY**, the **TO EAT VP**, and the **GO OUT TO EAT VP**. We stack the CF, AF, and PFL of S3.

Focus Data Structures after S4:

CF	MONEY
PFL	COLLEGE , the SPEND VP , ALMOST EVERY DAY , the TO EAT VP , and the GO OUT TO EAT VP .
AF	I
CF stack	MONEY , I , HOME
PFL stack	PFL of S3, followed by the PFL stack of S3

S5 contains a subordinate clause with a missing object. Our co-specification algorithm first considers the CF, **MONEY**, as the co-specifier of the omitted NP; semantics, syntax, and inferencing with discourse and general knowledge do not prevent this co-specification, so it is adopted. The co-specification algorithm is then applied to other NP’s. The Focus Tracking Algorithm chooses **MONEY** as the CF, since it is the co-specifier of an omitted non-agent NP occurring in the preferred clause of this sentence (i.e., the verb complement clause).

Focus Data Structures after S5:

²⁰If we were instead to split the sentence up, and treat each clause as a sentence, then the focus would shift away from **MONEY** when we process the second clause (which contradicts our intuition of what the focus is in this paragraph).

²¹The appropriateness of placing elements from both clauses in one PFL and ranking them according to clause membership will be further investigated. This construct (“X BECAUSE Y”) is further discussed in section 8.

CF	MONEY
PFL	NEED VP, MY PARENTS, HOME, GIVE VP
AF	I
CF stack	MONEY, MONEY, I, HOME
PFL stack	PFL of S4, followed by the PFL stack of S4

The next sentence of the text (S6) confirms MONEY as the CF of S5, thus giving some support to our expansion of the focus algorithm to handle omitted NP’s.

8 Discussion and Required Extensions

One of the major extensions needed in Sidner’s algorithm has to do with handling complex sentences. Based on a limited analysis of sample texts, we propose that we will compute the CF and PFL of a complex sentence based on a classification of sentence types. For instance, for a sentence of the form “X BECAUSE Y” or “BECAUSE Y, X”, we prefer the expected focus of the effect clause as CF, and order elements of the X clause on the PFL before elements of the Y clause. Analogous PFL orderings apply to other sentence types described here. For a sentence of the form “X CONJ Y”, where X and Y are sentences, and CONJ is “and”, “or”, or “but”, we prefer the expected focus of the Y clause. For a sentence of the form “IF X (THEN) Y”, we will prefer the expected focus of the THEN clause, while for “X, IF Y”, we will prefer the expected focus of the X clause. For a sentence with a verb complement, we will prefer the thematic positions of the verb complement before all other thematic positions (i.e., order all other thematic positions of the verb complement before thematic positions of the matrix sentence). Further study is needed to determine other preferences and actions (including how to further order elements on the PFL) for these and other sentence types. These preferences will likely depend on thematic roles and syntactic criteria (such as whether an element occurs in the clause that contains the expected CF).

We also need to explore AF calculation for complex sentences. At this point, we will pick the AF from the preferred clause, and put the agents in other clauses at the beginning of the PAFL, followed by all other non-agent (animate) NP’s.

The decisions about how these and other extensions should proceed have been or will be based on analysis of both standard written English and the written English of deaf people. The algorithm will be developed to match the intuitions of native English speakers as to how focus shifts.

A second difference between our algorithm and Sidner’s is that we stack the PFL’s as well as the CF’s. Some NP omissions we have analyzed require returning to a stacked PFL. It seems reasonable that stacking the PFL’s may be needed for processing standard English (and not just for our purposes). One reason we believe we need to stack PFL’s is it seems that sometimes, in complex sentences, focus revolves around the theme in one of the clauses, and later it returns to revolve around items in another clause. Further investigation may indicate that we need to add new data structures or enhance existing data structures to handle focus shifts related to these and other complex discourse patterns.

We should note that while we prefer the AF as the co-specifier of an omitted agent NP (recall our discussion of step 3, above), Sidner’s recency rule²² suggests that perhaps we should prefer a member of the PFL if it is the last constituent of the previous sentence (since a null argument seems similar to pronominal reference). However, our studies show that a rule analogous to the recency rule does not seem to be needed for resolving the co-specifier of an omitted NP. In addition, Carter [Car87] feels the recency rule leads to unreliable predictions for co-specifiers of pronouns.

²²Sidner’s recency rules prefers a member of the PFL which occurred as the last constituent in the last sentence as a co-specifier of a subject pronoun.

Thus, we do not expect to change our algorithm to reflect the recency rule. (We also suspect we will abandon the recency rule for resolving pronouns.)

The analysis given above for filling in the missing NP of S4 fills the NP based on the focus information from the previous sentence. Alternately, we can consider filling missing NP's in relative clauses with the topic of the main clause. This would be consistent with an analysis where the relative clause is assumed to be topicalized with a topic based on the main clause. We need to explore this alternative analysis for filling in NP's to determine which is more accurate or under which conditions each analysis should be applied.

Another task is to specify focus preferences among stacked PFL's and stacked CF's, perhaps taking thematic and syntactic information into consideration.

9 Example 2

Example:

“There are many things I like about NTID. They offer supporting services like interpreters and notetakers for mainstream classes which I had experiences through my public schools. Now NTID/RIT offers same thing that my school offered but only better supporting services. That is why I like about NTID.

(S1) But one thing worries me that most about NTID/RIT is financial problems.

(S2) I hope I could find some ways to solve __.”

9.1 Discussion and Required Extensions

An important question raised by this example is how to handle a paragraph-initial, but not discourse-initial, sentence. Do we want to treat it as discourse-initial, or as any other non-discourse-initial sentence? At this point, we suggest (based on analysis of samples) that if its sentence type fits a particular class of types, that we would use the preferences of the Discourse Initial Algorithm for calculating the CF and PFL to calculate the CF and PFL (in this sense, treat the sentence as discourse initial) and retain the CF and PFL stacks, pushing the last CF and PFL, (in this way, treating the sentence as not discourse-initial). Handling a paragraph-sentence in this manner allows for certain syntactic structures to play a more prominent role than they would otherwise. Two sentence types that would be included in this class of sentence types are mentioned by Sidner (p. 284, [Sid83]):

agent “There once was a prince who was changed into a frog.”

object “There was a tree which Sanchez had planted.”

We include sentences starting with “First,”, “Second,”, “Third,”, etc. in this class of sentences. We will explore whether other sentence types should be included in this class.

If a paragraph-initial sentence does not fall into this class of sentence types, we will treat it as any other non-discourse-initial sentence.

In this example, we will treat S1 as non-discourse-initial. First we use the co-specification algorithm to find co-specifiers of anaphora. Since S1 is a pseudo-cleft agent sentence, using the Focus Tracking Algorithm, we will pick the cleft agent as the CF. Thus, after S1, the CF is FINANCIAL PROBLEMS, the PFL contains THE ONE THING THAT WORRIES ME THE MOST ABOUT NTID/RIT, ME, and NTID/RIT. The AF is ME. (We stack the previous CF, AF, and PFL of the previous sentence.)

Focus Data Structures after S1:

CF FINANCIAL PROBLEMS
PFL THE ONE THING THAT WORRIES ME THE MOST ABOUT NTID/RIT, ME, and NTID/RIT
AF ME
CF stack the CF of the previous sentence, followed by the CF stack of the previous sentence
PFL stack the PFL of the previous sentence, followed by the PFL stack of the previous sentence.

We need to fill in a non-agent empty NP in S2. We first test the previous CF, **financial problems**. Since there is no reason to reject the previous CF as the referent of the empty NP, the algorithm fills the empty NP with the old CF. Thus, the empty NP is correctly filled by the algorithm.

10 Example 3

(S1) I think that Gallaudet College should require all deaf students to take speech and speechreading courses.

(S2) Therefore, they can improve their oral skills for their future use.

(S3) I am going to tell you that why the deaf student should take _”

10.1 Discussion and Required Extensions

This example illustrates an assumption discussed in section 5.3. The assumption was that if there is an omitted NP in a clause, all NP's that co-reference that NP in that clause are also omitted. As a result, we will reject any filling of an NP that makes it impossible to find the co-specifier of a full definite noun phrase in the same clause.

Under the assumption discussed above, the focusing algorithm functions as follows. After S1, by the Discourse Initial algorithm, the CF is **SPEECH AND SPEECH-READING COURSES** since that is the theme of the verb complement, and the PFL contains **DEAF STUDENTS**, the **TAKE SPEECH AND SPEECH READING COURSES VP**, **GALLAUDET COLLEGE**, and the **REQUIRE VP**. By the Actor Focus Algorithm, the AF is **DEAF STUDENTS** and the PAFL is **GALLAUDET COLLEGE**.

Focus Data Structures after S1:

CF SPEECH AND SPEECH-READING COURSES
PFL DEAF STUDENTS the TAKE SPEECH AND SPEECH READING COURSES VP,
 GALLAUDET COLLEGE, and the REQUIRE VP
AF DEAF STUDENTS
CF stack empty
PFL stack empty

Sidner's (third person agent pronoun) anaphora resolution algorithm would correctly predict actor ambiguity for "they". (It isn't clear whether the students will improve their abilities or the courses will.) We will assume our oracles will make the same predictions. If we assume "they" of S2 refers to **DEAF STUDENTS**, then, by the Focus Tracking Algorithm, since only members of the PFL are co-specified by anaphora, the CF becomes **DEAF STUDENTS**, and by the Potential Focus Algorithm, the PFL contains **THEIR ORAL SKILLS**, **THEIR FUTURE USE**, and the **CAN IMPROVE VP**, and the CF stack contains **SPEECH AND SPEECH READING COURSES**.

Focus Data Structures after S2 (“they” co-specifies DEAF STUDENTS):

CF	DEAF STUDENTS
PFL	THEIR ORAL SKILLS, THEIR FUTURE USE, and the CAN IMPROVE VP
AF	DEAF STUDENTS
CF stack	SPEECH AND SPEECH READING COURSES
PFL stack	DEAF STUDENTS the TAKE SPEECH AND SPEECH READING COURSES VP, GALLAUDET COLLEGE, and the REQUIRE VP

Then, for S3, the co-specification algorithm tries to fill the empty object NP with the previous CF (the first choice for an omitted non-agent NP). However, we reject this (DEAF STUDENTS) because “deaf students” of “... why the *deaf students* should...” can not co-specify with the filled omitted NP (based on our assumption that an omitted NP can not co-specify any non-empty NP in the same clause). Next, we try filling the NP with members of the PFL; semantics causes the rejection of THEIR ORAL SKILLS and THEIR FUTURE USE, and a VP can not fill an NP. So, we look at the focus stack, which contains SPEECH AND SPEECH READING COURSES. There is no reason to reject this and, thus, the missing NP is correctly filled by the algorithm.

If we assume “they” of S2 to refer to SPEECH AND SPEECH READING COURSES, then since “they” co-specifies the previous CF of S1, “their oral skills and “their future use” both co-specify DEAF STUDENTS, which is a member of the previous PFL, we must choose the new CF from among the previous CF and these PFL members. Since the anaphor co-specifying the CF is in agent position, we shift the focus to a member of the PFL co-specified by a non-agent. In this case, only one member of the PFL is co-specified by the remaining anaphora, so we select it (DEAF STUDENTS) as the CF. (We need to expand the algorithm to handle multiple anaphora co-specifying multiple members of the PFL.) The new PFL contains THEIR ORAL SKILLS, THEIR FUTURE USE, and the CAN IMPROVE VP. SPEECH AND SPEECH READING COURSES is pushed on the CF stack.

Focus Data Structures after S2 (“they” co-specifies SPEECH AND SPEECH READING COURSES):

CF	DEAF STUDENTS
PFL	THEIR ORAL SKILLS, THEIR FUTURE USE, and the CAN IMPROVE VP
AF	DEAF STUDENTS
CF stack	SPEECH AND SPEECH READING COURSES
PFL stack	DEAF STUDENTS the TAKE SPEECH AND SPEECH READING COURSES VP, GALLAUDET COLLEGE, and the REQUIRE VP

The analysis for filling the empty NP of S3 in this case (i.e., where “they” co-specifies SPEECH AND SPEECH READING COURSES), is the same as for the first case, since the contents of the focus data structures are the same after both case analyses of S2.

While both analyses yield the correct result for filling in the empty NP of S3, because of the unnecessary work that is required by considering all possibilities when we have an ambiguous pronoun, we anticipate that in the written English tutorial system, we will query the user as to which is the correct referent when we encounter an ambiguity condition.

11 Conclusions

We have discussed proposed extensions to Sidner’s algorithm to track local focus in the presence of illegally omitted NP’s, and to use the extended focusing algorithm to identify the intended co-specifiers of omitted NP’s. This strategy is reasonable since LT may lead a writer to use a rule of

discourse-oriented ASL which allows the omission of an NP that is the topic of a preceding sentence when writing sentence-oriented English.

The focus algorithm is potentially beneficial for correcting other errors in deaf writing. For example, it may be useful in identifying the intended referent when the writer has used a pronoun where a full definite description is required (to avoid ambiguity).

The major contribution of this thesis is the provision of a focusing algorithm for English that is more detailed and realistic with respect to English, especially for complex sentence types. Another contribution of this thesis is further documentation that there is transfer of discourse strategies in the language production of a second language learner. Additionally, it develops a methodology for capturing a particular class of errors expected to occur in the production of a sentence-oriented language being acquired by a speaker/signer of a discourse-oriented language.

12 Plan

We plan to study more text samples from ASL natives and from native speakers of English to test the proposed extensions and to identify extensions that will address other focusing questions which we've identified. In particular, we will look for other examples of NP omissions.

We will implement the focusing algorithm, assuming as input a GB-like syntactic parse tree and a semantic representation (as described in section 14.1) of the sentence. The co-specification confirmation will be done by an oracle since the required inferencing is beyond the scope of this work. The specification of the input requirements and output behavior of the oracle will be specified.

Finally, we will test our LT hypothesis and the extended focusing algorithm by testing whether the algorithm can correctly identify the omitted NP's in deaf writing samples which are different from the examples used in developing the algorithm.

13 Acknowledgments

We would like to thank John Albertini of the National Technical Institute for the Deaf (NTID), Bob McDonald of Gallaudet University, Lore Rosenthal of the Pennsylvania School for the Deaf, George Schellum (formerly) of the Margaret S. Sterck School, and MJ Bienvenu of the Bicultural Center for helping us gather writing samples.

A good part of our knowledge of ASL comes from discussions with ASL signers. We would like to thank our informants, April Nelson of Rosemont College, Don Ruble of Bloomsburg State College, and Jean Quillen and Carmine Salvato of the Pennsylvania School for the Deaf. We also thank Lore Rosenthal of the Pennsylvania School for the Deaf for interpreting for us. In addition, we want to thank the numerous people from deaf schools and organizations who have discussed this project with us.

We thank Julie Van Dyke for her implementation of the English grammar, and Jeff Reynar for his implementation of the error productions which will be used in the identification phase of the eventual overall system. We thank Karen Hamilton for the implementation of the database retrieval functions.

References

- [Bak80] C. Baker. Sentences in American Sign Language. In C. Baker and R. Battison, editors, *Sign Language and the Deaf Community*, pages 75–86. National Association of the Deaf, Silver Spring, MD, 1980.
- [BC80] C. Baker and D. Cokely. *American Sign Language: A Teacher's Resource Text on Grammar and Culture*. TJ Publishers, Silver Spring, MD, 1980.
- [BP78] C. Baker and C. Padden. Focusing on the non-manual components of American Sign Language. In P. Siple, editor, *Understanding Language through Sign Language Research*, pages 27–58. AP, New York, 1978.
- [BPB83] K. Bellman, H. Poizner, and U. Bellugi. Invariant characteristics of some morphological processes in American Sign Language. *Discourse Processes*, 6:199–223, 1983.
- [BS88] C. Baker-Shenk. Comparative linguistic analysis for interpreters. In D. Cokely, editor, *Sign Language Interpreter Training Curriculum*, pages 84–108. Fredericton, NB: University of New Brunswick, 1988.
- [Car87] David Carter. *Interpreting Anaphors in Natural Language Texts*. John Wiley and Sons, New York, 1987.
- [Dah86] Deborah A. Dahl. Focusing and reference resolution in PUNDIT. In *Proceedings of the 1986 National Conference on Artificial Intelligence*, pages 1083–1088, Philadelphia, PA, August 1986.
- [Fan83] Lou Fant. *The American Sign Language Phrase Book*. Contemporary Books, Inc., Chicago, 1983.
- [Fil68] C. J. Fillmore. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1–90, New York, 1968. Holt, Rinehart, and Winston.
- [Gas79] Susan Gass. Language transfer and universal grammatical relations. *Language Learning*, 29:327–344, 1979.
- [Gas84] S. Gass. A review of interlanguage syntax: Language transfer and language universals. *Language Learning*, 34(2):115–132, 1984.
- [Gas89] S. M. Gass. How do learners resolve linguistic conflicts? In S. M. Gass and J. Schachter, editors, *Linguistic Perspectives on Second Language Acquisition*, chapter 8, pages 183–199. Cambridge University Press, New York, 1989.
- [Gro77] Barbara Grosz. The representation and use of focus in dialogue understanding. Technical Report 151, SRI International, Menlo Park Ca., 1977.
- [GS83a] S. Gass and L. Selinker, editors. *Language Transfer in Language Learning*. Newbury House, Rowley, MA, 1983.
- [GS83b] S. M. Gass and L. Selinker. Introduction to section 3. In S. Gass and L. Selinker, editors, *Language Transfer in Language Learning*. Newbury House, Rowley, MA, 1983.
- [GS86] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July-August 1986.

- [Hak76] K. Hakuta. A case study of a Japanese child learning English as a second language. *Language Learning*, 26:321–51, 1976.
- [HS83] R. J. Hoffmeister and C. Shettle. Adaptations in communication made by deaf signers to different audience types. *discourse processes*, 6:259–274, 1983.
- [Hua84] C.-T. James Huang. On the distribution and reference of empty pronouns. *Linguistic Inquiry*, 15(4):531–574, Fall 1984.
- [Ing78] R. M. Ingram. Theme, rheme, topic and comment in the syntax of American Sign Language. *Sign Language Studies*, 20:193–218, Fall 1978.
- [KB79] E. S. Klima and U. Bellugi. *The Signs of Language*. Harvard University Press, Cambridge, MA, 1979.
- [KG83] J. Kegl and P. Gee. Narrative/story structure, pausing and American Sign Language. *Discourse Processes*, 6:243–258, 1983.
- [KK78] Richard R. Kretschmer Jr. and Laura W. Kretschmer. *Language Development and Intervention with the Hearing Impaired*. University Park Press, Baltimore, MD, 1978.
- [Kle77] H. Kleinmann. Avoidance behavior in adult second language acquisition. *Language Learning*, 27:93–108, 1977.
- [Lid80] Scott K. Liddell. *American Sign Language Syntax*. Mouton Publishers, 1980.
- [LM91] Diane C. Lillo-Martin. *Universal Grammar and American Sign Language*. Kluwer Academic Publishers, Boston, 1991.
- [MC88] Kathleen F. McCoy and Jeannette Cheng. Focus of attention: Constraining what can be said next. In C. L. Paris, W. R. Swartout, and W. C. Mann, editors, *Proceedings of the 4th International Workshop on Natural Language Generation*. Kluwer Academic Publishers, Boston, 1988. Santa Catalina Island, July.
- [McL87] R. McLaughlin. *Theories of Second-Language Acquisition*. Edward Arnold, London, 1987.
- [Odl89] T. Odlin. *Language Transfer*. Cambridge University Press, New York, 1989.
- [Pad81] C. Padden. Some arguments for syntactic patterning in American Sign Language. *Sign Language Studies*, 32:239–259, Fall 1981.
- [Pad82] C. Padden. *Interaction of Morphology and Syntax in American Sign Language*. PhD thesis, UCSD, 1982.
- [Pad88] C. Padden. *Interaction of Morphology and Syntax in American Sign Language*. Garland Publishing, Inc., New York, 1988.
- [PQ73] D. Power and S. Quigley. Deaf children’s acquisition of the passive voice. *Journal of Speech and Hearing Research*, 16:5–11, 1973.
- [QP84] S. P. Quigley and P. V. Paul. *Language and Deafness*. College-Hill Press, Inc., San Diego, 1984.

- [QPS77] S. P. Quigley, D. J. Power, and M. W. Steinkamp. The language structure of deaf children. *The Volta Review*, 79(80):72–84, February-March 1977.
- [QSW74] S. P. Quigley, N. L. Smith, and R. B. Wilbur. Comprehension of relativized sentences by deaf students. *Journal of Speech and Hearing Research*, 17:325–341, 1974.
- [QWM76] S. Quigley, R. Wilbur, and D. Montanelli. Complement structures in the language of deaf students. *Journal of Speech and Hearing Research*, 19:448–457, 1976.
- [RQP76] W. K. Russell, S. P. Quigley, and D.J. Power. *Linguistics and Deaf Children: Transformational Syntax and Its Application*. The Alexander Graham Bell Association for the Deaf, Inc, Washington, D.C., 1976.
- [Sac90] Oliver W. Sacks. *Seeing Voices*. University of California Press, Berkeley and Los Angeles, CA, 1990.
- [Sch82] J. Schumann. Simplification, transfer and relexification as aspects of pidginization and early second language acquisition. *Language Learning*, 33:337–66, 1982.
- [Sid79] Candace L. Sidner. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. PhD thesis, MIT, June 1979.
- [Sid83] Candace L. Sidner. Focusing in the comprehension of definite anaphora. In Robert C. Berwick and Michael Brady, editors, *Computational Models of Discourse*, chapter 5, pages 267–330. MIT Press, Cambridge, MA, 1983.
- [Sle82] D. Sleeman. Inferring (mal) rules from pupil’s protocols. *Proceedings of ECAI-82*, 9:160–164, 1982.
- [SR79] J. Schachter and W. E. Rutherford. Discourse function and language transfer. *Working Papers on Bilingualism*, 19:1–12, 1979.
- [Sto60] W. C. Stokoe, Jr. Sign Language structure. *Studies in Linguistics occasional papers*, (8), 1960.
- [Str88] Michael Strong. *Language Learning and Deafness*. Cambridge University Press, New York, 1988.
- [Sur91] Linda Z. Suri. Language transfer: A foundation for correcting the written English of ASL signers. Technical Report TR-91-19, Dept. of CIS, University of Delaware, 1991.
- [Wil77] R. B. Wilbur. An explanation of deaf children’s difficulty with certain syntactic structures of English. *The Volta Review*, 79(80):85–92, February-March 1977.
- [WS83] Ralph M. Weischedel and Norman K. Sondheimer. Meta-rules as a basis for processing ill-formed input. *American Journal of Computational Linguistics*, 9(3-4):161–176, 1983.
- [WVJ78] Ralph M. Weischedel, Wilfried M. Voge, and Mark James. An artificial intelligence approach to language instruction. *Artificial Intelligence*, 10:225–240, 1978.

14 Appendix A

All of the algorithms described in this thesis assume the following as input: a syntax tree and a semantic representation of the sentence which indicates which NP fills which thematic role [Fil68]. Further input specifications are given for each particular algorithm.

14.1 Discourse Initial Algorithm

The CF and PFL of a discourse-initial sentence are calculated slightly differently than for a non-discourse-initial sentence. Sidner referred to this algorithm as the Expected Focus Algorithm. We will use her algorithm, but refer to the output as the CF and PFL (rather than the expected focus and DEF) for simplicity. The algorithm relies on syntax in cases where the syntax is a strong indicator of focus (e.g., *There*-insertion sentences). If this is not the case, then the thematic roles of the sentence (giving preference to theme) are taken as the indicators of focus.

```
IF (the sentence is an is-a sentence)
THEN
    CF=the subject of the sentence.
ELSE IF (the sentence is a there-insertion sentence)
    THEN
        CF=object of the there-insertion sentence
    ELSE
        IF (THEME is a verb complement)
        THEN CF = THEME of verb complement;
            ELSE % THEME is not a verb complement
                CF = THEME;
PFL = all other thematic positions
    with the AGENT last,
    followed by the verb phrase.
IF the sentence has an agent
THEN AF = agent
ELSE retain AF;
PAFL = all NP's;
```

Extensions related to those discussed for the PFL in sections 7 and 8 will be needed to calculate the CF and PFL of a discourse-initial complex sentence.

14.2 Filling in a Missing NP (Co-specification Algorithm)

Missing non-agent NP:

```
Try to fill empty NP with CF
Try to fill empty NP with members of PFL
Try to fill empty NP with stacked CF's, and stacked PFL elements,
    under preferences yet to be determined. (For example, should
    one go through all CF's before PFL's or go down the CF and PFL stacks layer
    by layer?)
```

Missing agent NP:

IF trying to fill an agent NP in a simple sentence or in the first clause of a complex sentence

THEN

Try to fill empty NP with AF

Try to fill empty NP with PAFL

Try to fill empty NP with CF

Try to fill empty NP with stacked AF and PAFL

ELSE

IF trying to fill an NP in agent position in a complex sentence in other than the first clause

THEN

Try to fill empty NP with agent of the previous clause

Try to fill empty NP with AF

Try to fill empty NP with PAFL

Try to fill empty NP with CF

Try to fill empty NP with stacked AF and PAFL

(IF Discourse-Initial sentence, then fill with ‘I’)

(For anaphora which are non-empty, we rely on oracles to give us the co-specifiers.)

14.3 Proposed Focus Tracking Algorithm

Sidner mentions four sentence types which strongly mark focus and are usually not discourse initial (p. 284, [Sid83]):

pseudo-cleft agent “The one who ate the rutabagas was Henrietta.”

pseudo-cleft object “What Henrietta ate was the rutabagas.”

cleft agent “It was Henrietta who ate the rutabagas.”

cleft object “It was the rutabagas that Henrietta ate.”

In order to recognize the strong focus marking tendencies of the syntactic structures, the first thing the focusing algorithm does is test whether the sentence is of one of these types, and, if so, sets the CF accordingly.

IF cleft or pseudo-cleft

THEN IF the cleft item is not the previous CF, and some piece of the non-clefting item co-specifies with something in the focus data structure

THEN CF is the cleft item

ELSE the sentence is incoherent

The Focus Tracking algorithm is used to update the CF based on the focusing data structures and the co-specifications of the anaphora in the current sentence. The algorithm here is based on Sidner’s algorithm, and we have indicated which parts of the Focus Tracking Algorithm correspond to which steps of Sidner’s algorithm by including step numbers in the comments. We have omitted the steps corresponding to do-anaphora (since we do not handle verbal anaphora) and focus sets

(for clarity in the algorithm presentation; focus sets are rarely needed). Additions have been made to handle tracking focus in the presence of omitted NP's. Further modifications to handle complex sentence types must be made.

Input:

- CF (current focus) - the focus of the previous sentence
(Based on Discourse Initial Algorithm at start of sentence 2; otherwise the CF is determined by the last iteration of this algorithm.)
- PFL = Potential Focus List from the previous sentence.
(Based on Discourse Initial Algorithm if on sentence 2, based on PFL algorithm otherwise.)
- CF, PFL stacks - history of past CF's and PFL's; both empty on sentence 2
- information on anaphora in the sentence and their co-specifications with elements in focusing data structures .

Note: In what follows, the term “anaphor” includes an omitted NP.

```

Stack the old CF;
IF cleft or pseudo-cleft % Strong Syntactic indicators override
                        % usual rules based on thematics, syntax
                        % and focus history
THEN IF the cleft item is not the previous CF, and some piece of the
      non-clefting item co-specifies with something in the focus
      data structure
      THEN CF is the cleft item
      ELSE the sentence is incoherent
ELSE
  IF (there are multiple anaphora, at least one specifying the
      CF and at least one specifying something on the PFL)
  THEN BEGIN % step 3
    IF there was an omitted non-agent NP
      AND (the omitted non-agent co-specifies the CF or
           something on the PFL)
    THEN CF=omitted non-agent NP
    ELSE
      IF (there are anaphora co-specifying the CF and some
          members of the PFL)
        % Need to expand this part to handle multiple
        % anaphora co-specifying multiple members of PFL
      THEN IF (the cospecifier of the CF is a non-agent
              AND the cospecifier of the PFL is an agent)
            THEN
              retain the CF as focus;
            ELSE IF (the cospecifier of the CF is an agent
                    AND the cospecifier of the PFL is a non-agent)
                  THEN
                    make the new CF the old PFL element (for

```

```

        multiple PFL co-spec's, prefer
        pro's over full definite NP's, and consider
        PFL order (to be determined) if
        choice is still ambiguous
    ELSE IF (the CF co-specifies a non-agent
            AND the PFL co-specifies a non-agent)
    THEN IF only the PFL member is mentioned by a
            pronoun
            THEN make the CF the PFL member
            ELSE retain the CF;
END % step 3
ELSE IF (CF is co-specified by an anaphor, but no member of
        the PFL is co-specified by an anaphor)
    THEN BEGIN % step 4
            retain the CF as focus;
    END % step 4
    ELSE IF (anaphora cospecify members of the PFL, but no
            anaphor co-specifies the CF)
    THEN BEGIN % step 5
            IF (only one member of PFL is specified)
            THEN
                CF is that member of the PFL;
            ELSE
                choose CF in manner suggested
                by the ordering of the PFL (to be determined)
    END % step 5
    ELSE IF (there is an omitted NP in agent position)
            % since we are in this step, we know
            % no anaphor co-specifying the CF or
            % the PFL was found
    THEN CF=omitted NP;
    ELSE IF (the anaphora cospecify a member of
            the focus stack) % (but no anaphor
            % co-specifies the CF or a PFL member)
    THEN BEGIN % step 6
            move the CF to the stack member
            by popping the stack
    END % step 6
    ELSE % step 8
            IF ((no anaphora cospecifying any of CF,
                PFL, or focus stack)
                AND (CF can fill a non-obligatory
                    case OR the VP is related to
                    the CF by nominalization))
    THEN retain the CF;
            ELSE % step 10
                IF (no foci mentioned)
                THEN BEGIN

```

```

retain CF as focus;
for any unspecified pronoun, the
non-antecedent pronoun condition
holds
END % step 10

```

14.4 The PFL Algorithm – How to compute the PFL for a simple *non-discourse-initial* sentence

At the end of processing each non-discourse initial sentence, we compute the PFL. For any simple sentence, the potential focus list consists of a list of all elements in the knowledge network which are specified by NP's filling a thematic role²³, excluding an NP in agent position and excluding the NP which co-specifies the CF, followed by the verb phrase. This description fits that of Sidner's PFL algorithm.

We propose to explore whether we can further order elements of the PFL, based on thematic, syntactic or other criteria. One possibility is that we will order the elements by favoring elements filling obligatory roles over non-obligatory roles. We also plan to explore how (i.e., in what form) the VP should be included on the PFL. At this point, we believe that since all NP's that are related to the VP are already on the list, maybe all we need to do is put the verb on the list in order to handle nominalizations that serve as anaphora. We do not intend to handle VP anaphora.

In sections 7 and 8, we discussed how we need to extend this algorithm to handle complex sentences

14.5 AF algorithm

```

PROC calculate the AF;
BEGIN
  Stack the AF and PAFL;
  IF the sentence has an agent
  THEN AF = agent
  ELSE retain AF;
  PAFL = all NP's;
END

```

²³Throughout this work, when we write that the CF is an NP, or the PFL contains an NP, we mean that the data structure contains an element in the knowledge network which is specified by that NP.

15 Appendix B

- **Conjunctions:** 4
 - Omitted conjunction: 1
 - “He taught directed, for almost 30 years . . .”
 - Inappropriate conjunction: 3
 - “Other thing that I don’t like is some oral people talking with me without sign language but I can only understand in body language with oral.”
 - “*my classmate* is deaf and *some of them* can hear a few things.”
- **Prepositions:** 66
 - Omitted preposition: 26
 - “My brother like to go Castle Mall.”
 - “... the sign ‘ONE-DAY-PAST’ is glossed such words...”
 - Inappropriate preposition: 27
 - “My dolls are hanging *at* my wall.”
 - Extra Preposition: 13
 - “We help with each other with problems or anything.”
- **Determiners:** 63
 - Omitted determiner: 35
 - “Then we ate in bus.”
 - Inappropriate determiner: 9
 - “I will to build *the* sandcastle.”²⁴
 - Extra Determiner: 19
 - “*A* very little study was done...”
- **Incorrect Number on Noun:** 23
 - “... in *several language*.”
- **Incorrect Subject-Verb Agreement:** 11
 - “*My brother like* to go...”
- **Tense and Aspect:** 70
 - Dropped tense: 5
 - “The women’s dormitory *was* clean and *smell* good.”
 - Incorrect passive formation
 - “...they were not *permit* to enter due to their clothes.”
 - “Suppose *it was someone in your family that getting killed*.”
 - Incorrect BE/HAVE/DO auxiliary pairing
 - “They were both English teachers and I *do not heard* from them since I left to college.”
 - (“...*have not*...”)
 - Verb subcategorization problems
 - “I really *enjoyed to work* at my father’s office.”
 - “Third, Gallaudet will have more people interested *to enroll* Gallaudet University.”

²⁴Some examples are errors in the context in which they occurred, but the sentences appear correct in isolation.

- Problems related to use of “to”
“The boys *drive* car and *to listen* the music.”
- Extra, Incorrect or Omitted Modal : 2
“All persons guilty of drunk driving __ be sent to jail.” (should be sent) “I have more positive to do and __ handle myself.” (“can handle”)
“They *should* need to communicate more and meet more people to communicate each other.” (“They need”)
“I might need more time to find right people or my reputation *will* become bad.” (“might” - inappropriate)
- Other tense/aspect problems: 65
“I can go anywhere *without clean* my room...”
- **Mixing up English words or phrases which share a single ASL sign:** 12
“Third, living at home is *bored* and quiet.”
 - Omitted BE: 9
“Once the situation changes they __ different people.”
 - Lack of BE/HAVE distinction: 7
“... some birth controls *are* side-effect.”
- **Other Omitted Main Verbs:** 7
(Usually only with dummy subjects, if not be or have) “I enjoy with NTID student who is Deaf person and NTID staff and I like to talk with old NTID student because I like to hear about NTID’s History.” (“talking” or “being”)
“Better wait until I lived here more than one month.” (“It would be better...”)
- **Incorrect WH-phrase:** 4 “Now you can see *what* I compared these two of my teachers.”
(Correction, from context: “... *how* I compared...”)
- **Adjective Problems:** 13
 - Incorrect Adjective Choice: 3
“Especially, I do feel good to have here Because the food are the best than Gallaudet College.”
 - Incorrect Adjective Formation: 10
“it was very *complicate* to know where exactly is the bank.”
- **Incorrect Nominalization:** 5
“I, myself, will call the drunk driver a *murder* if he hit and killed my folks.”
“I have to learn alot of responsibles at NTID.”
- **Relative Clauses:** 14
 - Relative pronoun deletion: 4
“Then we go to see President from 1960 to 1963 __ is John Kennedy.” (“*who was John Kennedy.*”)
 - Resumptive pronoun: 1
“When I came to NTID for the first time, I met all my old friends that I didn’t expected them to come to same school.”
 - Other: 9

- **Pronouns: 12**

- Incorrect pronoun choice (including pleonastic): 7
“The students should have *them*...”
 (“them” refers to “birth control.”)
- Inappropriate pronoun use (where full definite descriptions are required): 4
“Fraternities and Sororities will see each other again like an old time. If Gallaudet should not allow Greek organization to continue, *they* will not cooperate each other since they do not know each other very well.”
- Lack of pronoun use (overuse of definite descriptions): 1
“My father hired me to run for my dad.”
“An airplane is better than driving a car. An airplane is very safe to go on a trip than driving a car. An airplane is faster than a car. An airplane can takes more people on the plane than less people in a car. A car is cheaper to go anywhere, but an airplane is more expensive to fly. The people can see many things happening in a car than flying on an airplane to see a plain sky.”

- **Pleonastic Pronoun Deletion: 10**

- Object: 5 “I loved _ here at Rochester Institute of technology because it was very beautiful place...”
- Subject: 5 “Better wait until I lived here more than one month.”²⁵
“The people are very friendly and interesting to get to know then who are from all over the U.S.”

- **Focus/Discourse Structuring**

- Problems: 49**

- Omission of focused element (subject: 4; object: 4): 8
“I hope I could find some ways to solve _.”
- Problems carrying over general/specific description strategies: 5
“Fraternities and Sororities here at XYZ DO provide social life. Some examples: parties; gettogethers; workplaces; IM; and sports.”
- Structuring problems with “because”: 8
“Only one thing that I don’t like NTID because of student always bothering me while I’m at dorm.”
- Ambiguous modifier attachment: 1
“There are many things I like about NTID. They offer supporting services like interpreters and notetakers for mainstream classes which I had experiences through my public schools. Now NTID/RIT offers same thing that my school offered but only better supporting services.”
(*Ambiguous as to what student had experience with.*)
- Other (possibly related to carry-over of topic-comment strategies): 27

- **Redundancy Problems: 2**

- “I still feel thankful for coming to NTID instead of Gallaudet the main reason why I stay here, is the warm feeling everyone have toward the others.”

²⁵Often more than one correction is possible. For example, here the correction could be “It is better to wait until ...” or “I had better wait...”

- **Not Enough Sentence Breaks:** 6
- **Other:**104 (23% of errors in database)