Speech Repairs in the DIME corpus

Iván Moreno[a,b], Luis A. Pineda[a]

[a] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS, UNAM)
imoreno@turing.iimas.unam.mx, luis@leibniz.iimas.unam.mx
[b] Facultad de Ingeniería, UNAM

Abstract. In this paper the analysis of speech disfluencies and repairs in a task oriented spoken corpus in Spanish, the DIME Corpus DIME [11], and its relation to dialogue segmentation and discourse markers is presented. A method to detect and correct the most common speech disfluencies for speech recognition in this corpus is also presented.

1. Introduction

One of the main aims of natural language processing and speech recognition is to develop computational systems able to engage in task oriented natural language conversations. At the current state of the technology it is possible to construct dialogue systems focused on domain specific tasks in which simple and grammatical language is used. Current speech recognition systems provide a set of weighed hypothesis of what the speaker is supposed to have said and, usually, the hypothesis with the highest weigh is taken for further processing steps; in particular, the parser finds the syntactic structure and semantic representation of this textual input. In the ideal case, sentences are meaningful and well-formed and the parsing process can proceed in the standard pipe-line architecture; however, spontaneous language, commonly used in conversations, exhibits interjections, pauses, repetitions, etc., and also ungrammatical language, that must be dealt with in order to construct useful systems. These spontaneous speech phenomena are called *speech disfluencies*. In order to process the spoken input disfluencies must be *corrected* and this is usually done within the same elocution, as exemplified by the following elocution taken from corpus.

<sil> *la estufa pegar vamos a quitar* <sil> a <sil> *a intercambiar vamos a poner este fregadero esto* <sil> *de este lado de acá y la estufa de este lado de acá* <sil>

<sil> *the stove place let´s take off* <sil> to <sil> *to interchange let´s put this sink this* <sil> *at this side of here and the stove at this side of here* <sil>

The process of obtaining the intended elocution from the elocution containing disfluencies and correct is also called as elocution's correction process or simply, correction process. Speech disfluencies and repairs do not obey grammatical rules and robust speech recognition systems must have a model to detect and correct them in order to facilitate or even make possible the parsing process; in this paper a case study of disfluencies and repairs appearing in the corpus DIME is presented; also the construction of decision trees to detect disfluencies and repairs is presented, and a simple algorithm to correct speech repetitions, the most common kind of disfluency, is presented too.

2. Phenomena of the spontaneous speech

Speech disfluencies and repairs are related to dialogue segmentation and the presence of discourse markers. Unlike written language where the sentence is a well-defined and understood notion, there is not a natural unit of speech. In order to understand and analyze spoken language, the continuous flux of discourse needs to be divided in manageable units that express basic intentions and correspond roughly to units of intonation and meaning (i.e. speech acts). These units are commonly referred to as *utterances* and the process of dividing the discourse into utterances is referred to as segmentation. The segmentation process is aided by words that mark the boundaries and relations between utterances and these words are referred to as *discourse markers*. Discourse markers also help to identify and correct speech disfluencies. In the rest of this section the notions of segmentation, repairs and discourse markers, as well as their relation, are further elaborated.

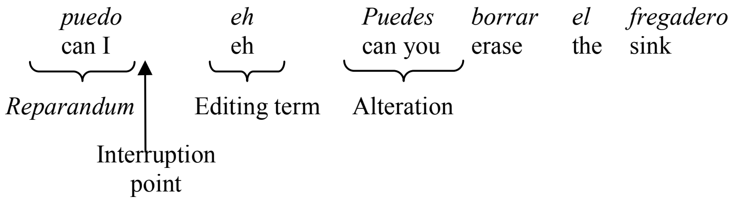
2.1 Segmentation

Human conversation is structured in terms of turns. In each turn a conversational participant expresses one or more intentions, or makes contributions to the effect that his or her beliefs and intentions are understood as intended, maintaining in this way the conversational agreement or common ground; however, the information expressed in each turn may be too large to be considered a natural unit of understanding, and each turn may be divided into one or more basic units or utterances. The proper segmentation of the hearer is fundamental to understand and proceed with the conversation, and also for discourse analysis, as will be seen below.

2.2 Speech repairs

Speakers make conversational contributions incrementally with the purpose to express intentions; however, it often happens that the expression of an intention process at the same time that the corresponding planning process and speakers may need to review and correct what they have already said. This kind of disfluencies interrupts the normal intention of utterances, and may contain pauses or discourse markers that signal the disfluency and the corresponding repair. Repairs can have different structures; for instance, what has been said can be abandoned completely, to start the idea afresh; speakers can also repeat some words to repair the utterance, or simply introduce a pause, perhaps filled with a word, like an interjection, to have some time to plan how to express the idea. All this forms of disfluenicies and their corresponding correction are referred to as *speech repairs*.

Speech repairs have a standard structure [10]; this structure contains three spoken components and one pitch element. The spoken components are called *reparandum*, *editing term* and *alteration*, and the pitch element is the time at which the disfluency is realized, and it is known as *interruption point*. Next, the standard structure of a repair is illustrated:



puedo *eh* *Puedes* *borrar* *el* *fregadero*

can I eh can you erase the sink

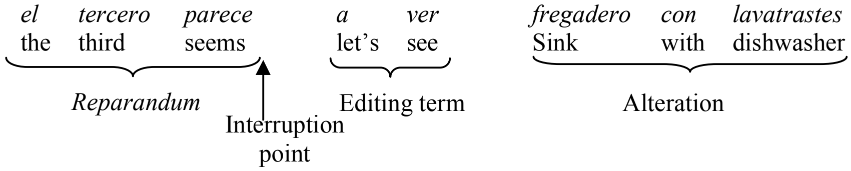
Reparandum Editing term Alteration

Interruption point

- *Reparandum*: the part of the utterance which the speaker wants to correct.
- *Interruption point*: the time point at which the disfluency is realizaed, with the corresponding distortion of the normal intonation pattern of the utterance, and it appears at the end of the *reparandum*.
- *Editing term*: a word or a phrase, with a predictable meaning, that is used to fill the pause needed to plan what will be said next; examples editing terms are interjections like *ah*, *mm* or *eh* and also some idiomatic phrases such as *es decir*/this is, *bueno*/well or *perdón*/pardon me
- *Alteration*: the part of the utterance that expresses the right idea and replaces the *reparandum*.

According to the relation between the *reparandum* and the alteration Speech repairs can be classified in three types [10]: *fresh starts, modification repairs* and *abridge repairs*, as follows:

- *Fresh Start*: the speaker simply abandons what he just said; there is not relation between *reparandum* and alteration.



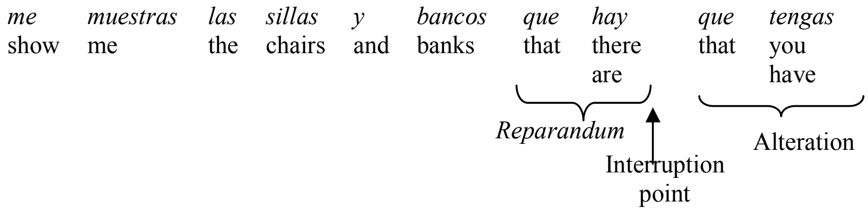
el *tercero* *parece* *a* *ver* *fregadero* *con* *lavatrastes*

the third seems let's see Sink with dishwasher

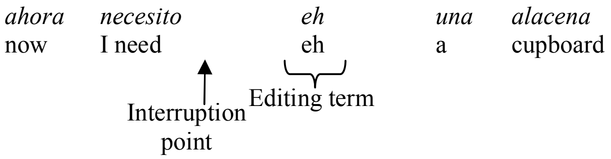
Reparandum Editing term Alteration

Interruption point

- *Modification Repair*: repairs with a non – empty *reparandum*.; most repairs are of this kind. A particular type of modification is the repetition repair, in which one or more words are shared by the *reparandum* and the alteration.


me muestras las sillas y bancos que hay que tengas
show me the chairs and banks that there are that you have
Reparandum
Interruption point
Alteration

- *Abridged Repair*: repairs without a *reparandum* which only present one or more editing terms.



2.3 Discourse markers

Discourse markers are words that have a discourse function instead of their usual lexical category in well-formed sentences, if there is such; words such as *bueno*/well, when it marks that a proposition has been accepted, or as when one answers the telephone; *entonces*/then, when this words establishes a causal relation between two propositions, and *ahora*/now, when the intentions is to mark that a new topic will be addressed, are examples of discourse markers. In such contexts these words do not function as adjectives or adverbs, as they become markers precisely when they abandon such a syntactic function; their purpose is to provide the listener with enough information about the structure of the dialogue and to guide the inference that has to be made to make the communication successful ([4], [5], [8], [9]).

2.4 Relations between phenomena

The phenomena of segmentation, repairs and discourse markers are highly inter-related. Discourse markers together with intonation are important cues for segmentation; discourse markers can also help to detect repairs, because the editing term is generally formed by this kind of markers. In particular, the abridge repair is characterized precisely by the appearance of a discourse marker, normally an interjection. The relation between segmentation and repairs is also complex, as the presence of an interruption point may be confused with an intonation boundary; this problem appears with fresh starts where the alteration can often be taken as a full utterance in itself.

3. The empirical study

In the present investigation the DIME corpus [11] was used as empirical base to study disfluencies and repairs. In addition to the tagging levels of this corpus, three extra levels were tagged: (1) speech repairs, (2) Part–Of–Speech (POS) and (3) identification of discourse markers. The word tagging level of the DIME corpus was used as a reference for these three new tagging levels; in addition the *break indices* level of the DIME Corpus, based in the ToBI [1] intonation tagging scheme was used.

3.1 Speech repairs level

This tagging level is formed by three sub–levels:

- Structure: A time aligned tagging of *reparadum*, editing term and alteration are marked in this level.
- Type: The type of speech repair (e.g. fresh start, modification or abridge)
- Repair's relations: This level codifies the relations between the words in the different parts of the repair's structure. This level is based on Heeman [9]. The tag set for this level is shown in Table 1.

Tag	Description
m*i*	Marks that two words are the same
r*i*	Marks that a word replaces another
x*r*	Marks that a word is deleted or inserted
p*i*	Marks a multi-word correspondence, such as the replacement of a pronoun by a longer description
srr<	Marks the onset of the *reparandum* of a fresh start
et	Marks an editing term

Table 1: Speech repair tag set

3.2 Part–Of–Speech (POS) level

In this level the lexical category of all words in the utterance are stated. The tag set for this level is based on the analysis of one dialogue of the DIME corpus, and also on using different proposals previously made in the literature both for English and Spanish ([2], [6], [7], [9]). The final tag set for this level is shown in Table 2.

Tag	Description
N	Noun
V	Verb
VAM	Auxiliary – Modal Verb
VC	Clitic Verb
A	Adjective

AD	Demonstrative Adjective
TD	Definite Article
TI	Indefinite Article
R	Adverb
RI	Interrogative Adverb
RR	Relative Adverb
RN	Negation Adverb
RA	Affirmation Adverb
P	Pronoun
PD	Demonstrative Pronoun
PR	Relative Pronoun
PI	Interrogative Pronoun
PC	Clitic Pronoun
S	Preposition
C	Conjunction

Table 2: Part-Of-Speech tag set

For instance: *Así/R está*/V *bien*/A? (Is this okay?)

3.3 Discourse markers level

As was mentioned, discourse markers are words in which a discursive function predominates over their usual syntactic function; on the basis of this consideration the tag of a discourse marker is formed with the tag of the normal lexical category of the word prefixed with MD (***M**arcador del **D**iscurso*). For instane: *ahora*/MDR *ponme la estufa* (Now, put the stove (for me)).

In addition, three new tags for words that do not have a lexical category were also included as shown in Table 3

Tags	**Description**
MDI	Interjection
MDK	Acknowledgment
MDeste	*este*

Table 3: Extra Tags for the discourse markers level

3.4 The tagging task

In the present investigation 8 dialogues of the DIME Corpus were tagged in these three levels; in this exercise 1105 utterance were tagged, out of which 105 presented a repair. Although the speech disfluencies are less than 10% of the data, repairs present characteristic patterns that can be used for the detection and correction task.

4. Detection of repairs

The empirical data was analyzed in order to find useful variables for the detection and correction of repairs. From this exercise four useful variables were found; these are utterances' duration, number of words, presence of a silence and the type of the dialogue act expressed by the utterance. These variables permitted to identify a detection strategy based on the construction of decision trees.

4.1 Detection variables

A basic intuition is that utterance with repairs should last longer that the corresponding utterance without a repair. This is corroborated in Figure 1 were it can seen that 77% of the tagged utterances last between 0 and 2000 milliseconds and only 1% of these utterance have a repair. On the other hand, a very large percentage of the remaining 23% presents one or more repairs.

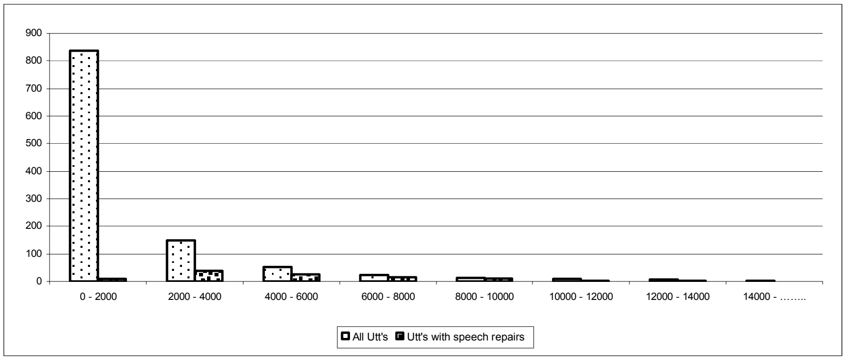


Figure 1: Speech repairs according to duration of the utterance

The second intuition is that an utterance with a repair has a larger number of words than the corresponding utterance without the repairs. This is verified in Figure 2, where utterances are classified in three classes according to their number of words. Region R1 contains utterances with 6 or less words; utterances in R2 contain between 7 and 15 words, and utterances in R3 have more than 15 words. As expected, 79% of all utterances are in R1, but only 2.34% of these have a repair; the critical region is R2 as it has 18.52 % of the utterances, and 30.69% of these have one or more repairs.

Finally, R3 has 3.02% of the utterances and 70% of these have a repair. A further analysis showed that the media of the time duration of utterances with repairs in R2 is longer than the media of the time duration of utterances without repairs in this region for utterances with the same number of words.

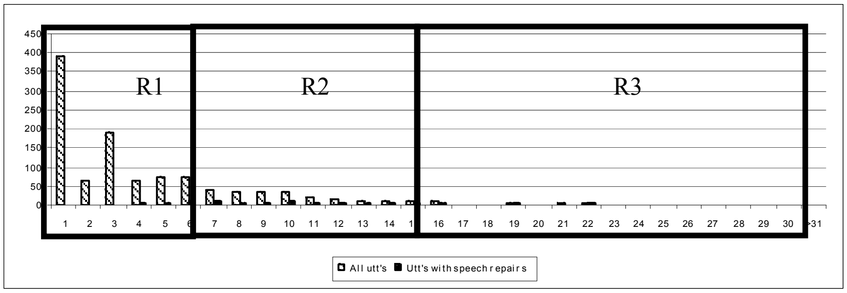


Figure 2: Speech repairs according the number of words in the utterance

The third intuition is that utterances with repairs should have a silence, as it is common that after the interruption point the speaker needs some time to re-elaborate the utterance. This is also verified by the data as 86% of utterances with at least one speech repair present a silence; in addition, a silence usually increases the utterance duration.

Finally, it was observed that speech disfluencies are related to the dialogue act expressed by the utterance; in particular, 64% of the repairs are action directives and 30% are affirms; the intuition behind this observation is that in the case of these two dialogue act types, the speaker is planning along the elocution of the utterance, while other dialogue acts may have a more reactive character.

4.2 Speech repairs detection

The four variables identified above suggested a detection strategy based on the construction of a decision tree. For this purpose utterance were classified using CART[1] style decision trees generated with the WEKA[2] tool.

For the construction of the decision tree 105 utterances with and 105 without repairs were taken. The same number of utterances with and without repairs was taken from each dialogue. This strategy helped to balance the process. The resulting decision tree is shown in Figure 3 and the statistics can be observed in Table 4.

[1] http://www.salford-systems.com/112.php

[2] http://www.cs.waikato.ac.nz/ml/weka/

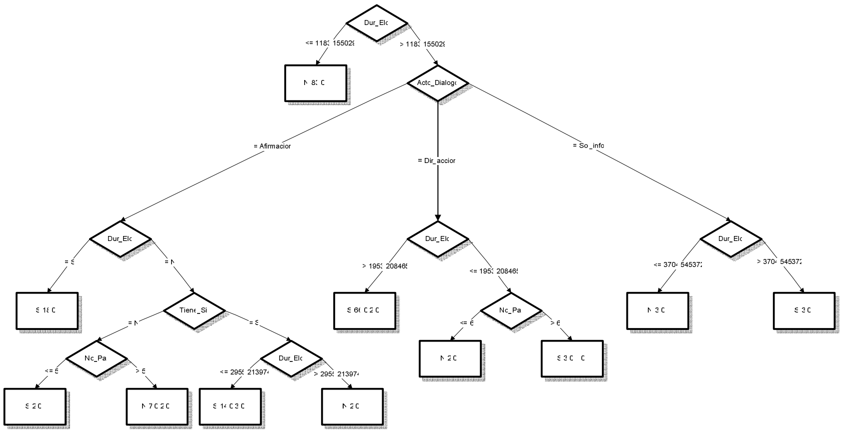


Figure 3: Decision tree to detect a speech repair

Correctly Classified Instances: 86.73%
Kappa: 0.73

	Precision	Recall
Don’t have repair	0.94	0.76
Have repair	0.82	0.95

Table 4: Speech repairs detections statistics

As expected, the main classifying attribute was utterance’s duration and the second most prominent one was dialogue act type. As the figures in Table 4 show, the classification task is very satisfactory both in terms of precision and recall.

5. Correction strategy

On the basis of the data analysis a simple strategy for the correction of the most common type of repair in the DIME Corpus, the repetition repair, was identified. In the current tagged data almost 79% of the utterances with one or more repairs are modification repairs; also, almost 77 % of these repairs are repetition repairs. This kind of repairs present, in addition, a simple and regular structure, and a simple heuristics to correct this kind of repairs was identified.

For the definition of the heuristics correction rule two variables were considered: the actual repetition of words and the distance between repeated words, measured in the number of words in between. The intuition is that when a word is repeated in a repair the repeated word appears immediately or in a close proximity of the repeated one. This was verified in the data as about 71% of the utterances with a modification repair present the repetition immediately (i.e. distance = 0) or with only one word in between (i.e. distance = 1); in addition, about 90% of the utterances present a repetition of word with a distance between 0 and 2. This is, most repetitions have a distance of less than 2. It is important to consider that common types of words, like articles and prepositions, are repeated in most utterance regardless whether there is a repair, but repetitions in utterances without repair usually belong to different constituents (e.g. subject and object) and the distance between the repeated words is almost all the time larger than two.

5.1 Correction algorithm

For the implementation of the detection and correction phase a pipe-line architecture was used; first, all utterances are classified through the decision tree, and those classified as positive are passed through the correction algorithm. The correction algorithm is illustrated next:

1. The input is the utterance produced by the Speech Recognition System; the words are indexed from 1 to N:

eh	*igual*	*con*	*la*	*con*	*la*	*estufa*
eh	same	with	the	with	the	stove

2. The repeated words are identified, and the distance between them is associated to the first instance of the repeated word:

eh	*igual*	*con*	*la*	*con*	*la*	*estufa*
eh	same	with	the	with	the	stove
		1	1			

3. If there are repeated sequences are identified and abstracted as units (i.e. the repetition chunk). In addition, distance between chunks is computed, and the value is associated to the first instance of the chunk.

eh	*igual*	*con*	*la*	*con*	*la*	*estufa*
eh	same	with	the	with	the	stove
		0				

The units (i.e. words or chunks) with distance less or equals than 2 are removed and the remaining units are attached to the remaining instance of the repeated unit as shown below:

eh	*igual*	*con*	*la*	*con*	*la*	*estufa*
eh	same	with	the	with	the	stove
		0				
eh	*igual*	*con*	*la*	*estufa*		

eh	same	with	the	stove

4. Else (i.e. there are no repetition sequences) if distance is less or equal than 2 remove the words within the distance from the first instance of the repeated word, and also remove the second instance of the repeated word.

entonces	*el*	*primero*	*el*	*tercero*
then	the	first	the	third
	1			
	1			
entonces	*el*	*tercero*		
the	the	third		

The algorithm was tested with the available data and the results are shown in Table 5, as follows:

	Was corrected	Was not
Should be corrected	55%	18%
Should not	4%	23%

Table 5: Speech repairs correction statistics

Table 5 shows that 78% of the utterances were correctly processed and only 22% were handled inappropriately by the heuristics. In particular, the decision classifies all type of repairs, and most fresh starts and abridge repairs where handle correctly by the method. On the other hand, out all repetition repairs 75% were handled correctly by the heuristics and only 25% of these were ignored or badly handle by the method. This can be considered a very promising result.

6. Conclusions

The phenomenon of speech disfluencies is very complex but it has to be faced directly in the construction of speech recognition systems. Heeman [9] provides a very complex method to handle this phenomenon through the definition of multidimensional language models; however, the present study shows that a simple detection strategy paired with a heuristics to correct the most frequent kind of repair can be very effective in the solution of this problem. The present is a preliminary experiment, and we hope that a larger amount of data may be useful to improve the classification rate, to distinguish different kinds of repairs, and to identify specific heuristics to deal with correct other kinds of disfluencies.

In addition additional tagging levels currently available in the DIME Corpus, such as a tonal analysis using INSINT [3] model, or the duration final vowel or consonant extensions can indicate the end of the *reparandum*, and this information can also be very useful for the detection of repairs and their kinds. It is also possible to consider specific discourse markers for the identification and correction of abridged repairs, and it may be possible to do this correction on the fly.

Acknowledgment

We would like to thank support of all members of the DIME project; their tagging work made possible the accomplishment of this work; we would like to mention Ivonne López and Hayde Castellanos for the great help during the POS tag set level definition and Sergio Coria during the decision trees' construction. Also, we thank useful comment and suggestins from James Allen at Rochester University and Joaquím Llisterrí at Universidad Autónoma de Barcelona have given us. The theory and experiments reported in this paper were developed in the context of the DIME-II project, with support of NSF/CONACyT grant 39380-A.

References

[1] Beckman, M., Diaz Campos, M., Tevis and J, Morgan, T. (2000). Intonation across Spanish, in the Tones and Break Indices framework, Forbus (14), pp. 9 – 36.

[2] D. Farwell, S Helmreich & M. Casper. SPOST: a Spanish Part-of-Speech Targger. http://crl.nmsu.edu/Publications/farwell/far_etal95.html

[3] Hirst, Daniel., Di Cristo, A., and Espesser, R. Levels of representation and levels of analysis for the description of intonation systems, In M. Horne (ed) Prosody: Theory and Experiment, Kluwer-Dordrecht, 2000

[4] http://es.wikipedia.org/wiki/Marcadores_del_discurso

[5] Julian Hirschberg y Diane Litman, Empirical Studies on the Disambiguation of Cue Phrases. Computational Linguistics, 19(3): pp. 501 – 530, 1993.

[6] M. Civit & M. A. Martí. Design Principles for Spanish Treebank. En proceedings of the First Workshop on Treebanks and Linguistics Theories (TLT2002), pp. 61 – 77, September 2002.

[7] M. Civit. Criterios de Etiquetación y Desambiguación Morfosintáctica de Corpus en Español. PhD thesis. Universidad de Barcelona, 2003.

[8] Maria Vittoria Calvi y Giovanna Mapelli, Los marcadores bueno, pues, en fin, en los diccionarios de español e italiano. Artifara, n. 4, sezione Monographica, http://www.artifara.com/rivista4/testi/marcadores.asp

[9] P. A. Heeman. Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog. PhD thesis. Rochester University, 1997.

[10] Peter A. Heeman, Kyung-ho Loken-Kim, y James F. Allen. Combining the Detection and Correction of Speech Repairs. Proceedings of 4rd International Conference on Spoken Language Processing (ICSLP-96), pp. 358–361, Philadephia, October 1996, also appeared in International Symposium on Spoken Dialogue, 1996, pp. 133-136.

[11] Villaseñor, L., Massé, A. & Pineda, L. A. (2001). The DIME Corpus, Memories 3º. Encuentro Internacional de Ciencias de la Computación ENC01, Volume II, C. Zozaya, M. Mejía, P. Noriega y A. Sánchez (eds.), SMCC, Aguascalientes, Ags. México, September, 2001.