Winner of the 1999 Presidents' Pre-Doctoral Prize

Presidents' Pre-Doctoral Prize

The Presidents' Pre-Doctoral Prize is awarded annually to the paper judged to make the greatest contribution to linguistic knowledge presented by an author who has not yet completed a doctorate. The judging panel consists of the current LACUS *President and Vice-President along with all past presidents in attendance at the meeting.*

COHESION AS A MEASURE IN GENERIC ANALYSIS

Maite Taboada
University of Alberta, Edmonton

Cohesion, a property of any successful text, is also present in spoken language. Speakers relate their utterances to previous ones through the use of cohesive relations; a cohesive tie is established. Cohesive ties enter into cohesive chains, which run throughout a text, revealing how different parts of a text are related to each other.

In this study I describe the results of an analysis of cohesive relations in a bilingual (English and Spanish) corpus of task-oriented dialogues. The two research questions I would like to address relate, first, to the relationship between the generic stages of the conversations and the boundaries of cohesive chains. The dialogues, as instances of a scheduling genre, proceed in clearly defined stages. I will show how cohesive chains are clearly related to stages in the dialogue. The second research question concerns measures of cohesive harmony in speech. The dialogues, although perfectly functional, seem to contain very low cohesive harmony. I suggest that different measures of cohesive harmony are necessary for different genres.

1. the corpus. The corpus used was a total of 60 conversations between dyads of two native speakers each, 30 in English and 30 in Spanish. For each language, the conversations are broken down in ten female-female, ten male-male, and ten female-male dialogues. The dialogues were collected by the Interactive Systems Laboratory of Carnegie Mellon University as part of janus, a large speech-to-speech machine translation project[1]. The dialogues are transcribed including all human and non-human noises, intonation, and divided into semantically independent units, or Semantic Dialogue Units (sdus). Sdus are clauses for the most part. Sometimes sdus are constituted by phrases or words that were intonationally separated from the rest of the turn. The translation of Spanish examples is a free translation, rendered one clause at a time[2].

The conversations are considered instances of a genre, as a purposeful, staged, goal-oriented activity (Martin 1985), which we might denominate *scheduling genre*. They are instances of talk produced for a very specific purpose, that of setting up an appointment. As a result of what we could call their practical purpose, the conversations are staged in particular ways: one speaker proposes a meeting, perhaps also a time; the other speaker replies either with a different time or with their availability for the time proposed. The conversation continues until a day and a time have been set. As a result of their social function, the conversations usually have Opening and Closing stages, and polite devices that will avoid face-threatening acts

to the other speaker. Given their task-oriented character, we can expect them to show a structure more uniform and less deviant from an idealized model than, for instance, casual conversation. The task usually has a clear structure, consisting of steps that have to be taken in a certain sequence in order to successfully complete the task.

There are three main stages in the conversations: Opening, Task Performance and Closing. The Openings and Closings resemble those present in telephone conversations. The speakers begin the conversation with brief greetings or vocatives, and close it by repeating the date agreed upon and signaling leave-taking. The Task-Performance stage is the main body of the dialogues, and it contains two other types of stages: Date Proposal and Place Proposal. The Date-Proposal stage is initiated when one speaker proposes a date, and it continues until the date has been accepted or rejected. If the date is rejected, one of the speakers will initiate a new Date-Proposal stage. The dialogues may, then, contain more than one Date-Proposal stage. Place-Proposal stages are optional; they occur when an appropriate locale for the meeting is arranged.

2. A BRIEF INTRODUCTION TO COHESION. According to Halliday and Hasan (1976), the property of being a text is called texture. Texture is what distinguishes a text from a non-text, and it is derived from the fact that the text functions as a unity with respect to its environment. Texture is realized in relations existing between parts of a text. Let us look at one example.

(1) …Would you like to meet possibly, between the seventh to the tenth? Anytime during those days would be fine.

In this example, *those days* refers to *between the seventh to the tenth*. There is a relation between those two phrases that makes the two sentences become a text, because they hang together as one unit. This relation is a *cohesive relation*, and the pair of related items is a *cohesive tie*. (Obviously, no single item is cohesive in itself. Although I will be referring to *those days*, and categorizing it as a particular type of cohesion, we should always bear in mind that cohesiveness is established through the relation between the two items, not by one item in isolation.)

The meaning of this relation is that the two items refer to the same thing; they share corefentiality. Identity of reference, as we shall see, is not the only type of cohesive relation, there also exist relations of similarity. But in a more general sense, cohesion occurs when the interpretation of some element in the discourse depends on the interpretation of another one, whether preceding or following. In our example above, *those days* presupposes *between the seventh to the tenth*. We need to refer to the latter in order to resolve the presupposition. The fact that it is resolved successfully establishes the cohesive relation between the two clauses.

There are, in fact, two different types of tie between *those days* and its presupposed element. *Those* establishes a relation of demonstrative reference, whereas *days*

is a general word that subsumes specific instances of days. Cohesive relations are classified as: reference, substitution, ellipsis, conjunction and lexical cohesion. The first three types are usually referred to as anaphoric relations. For the analysis described here, I excluded conjunction, because I believe that it captures slightly different relations—relations among clauses, rather than among referents in the text. Elsewhere (Taboada forthcoming) I performed an analysis of conjunctive relations in this corpus, through Rhetorical Structure Theory (Mann & Thompson 1988). The types of cohesion considered for the present analysis are listed in Table 1.

Cohesive ties are established between elements in the text (endophoric reference), not with elements that have their referent outside the text (exophoric reference). In addition, the relationship in a tie can be measured in terms of the distance between its components. The relationship might be immediate (the cohesive element refers to an immediately preceding one), remote (the referent is more than one clause away) or it can be mediated, where the ultimate referent is a few clauses earlier in the preceding discourse, but it has been recaptured in some other element. A cohesive element could also have a cataphoric relationship to some other element in the discourse to follow.

The concept of mediated ties brings us to the next important concept in cohesion, that of cohesive chains. If two ties are mediated by a third intervening one, then the three of them enter into a cohesive chain. Let us look at Example (2), from the Spanish corpus. One of the cohesive chains running through these two turns is one that relates to food. The speakers have decided on a lunch meeting, and thus *restaurante*, *comida* and *estomaguito* can all be related to this topic. The chain begins earlier on in the conversation, but here we can see part of it with six links (**bolded**). If there was only one preceding item, the chain would then be seven links long. All of the links are instances of lexical cohesion: repetition of the same item and collocation. Most texts contain more than one chain: in the example we can see there is another chain relating meeting times.

(2) **s1**: Okay, está, acordamos entonces que sea martes, me parece mejor a las doce y treinta, así yo termino clases a las doce, y alcanzo a llegar al **restaurante** a las doce y treinta *pause*. Lo que me parece más bien que definamos es a dónde quieres ir? Tú dime qué qué tipo de **comida** quieres para el martes. O si tu religión te impide /begin_laugh/ algún tipo de comida especial, entonces evitamos ese tipo /end_laugh/ de **restaurantes**.

s2: Correcto Mónica. Nos vemos entonces ese día a las doce y treinta. /eh/ Como te decía yo pasaré con Len, y no tengo ningún problema con la **comida**, como buen pobre y colombiano que soy *pause* tengo acceso a cualquier /begin_laugh/ tipo de **comida** /end_laugh/. Mi **estomaguito** me lo permite. Gracias mija.

(2) **S1**: (Translation) Okay, it's, we agree then that it's Tuesday. I think it's better at twelve thirty, that way I finish my classes at twelve, and I'll manage to get to the restaurant at twelve thirty *pause*. What I'd like for us to define is where you want to go? You tell me what what kind of food you want for Tuesday. Or whether your religion forbids any type of special food, then we'll avoid that type of restaurant.

S2: (Translation) Right Monica. We'll meet then on that day at twelve thirty. As as I was saying I'll come with Len, and I have no problem with the food, being a good poor and Colombian as I am *pause* I have access to any /begin_laugh/ type of food /end_laugh/. My stomach allows it. Thank you dear.

Cohesive chains and chain interaction are the most interesting constructs in describing cohesion and, ultimately, coherence in a text. Many scholars have pointed out that cohesion and coherence are not all-or-nothing categories, but a matter of degree. Parsons (1996) states that, in any given text type, there is a gradation dependent on the extent to which a text relies on cohesion to provide coherence. He takes as starting point the work of Hasan (1984), and Halliday and Hasan (1985) in the development of the concept of *cohesive harmony*. Hasan defined a chain as a set of items, each of which is related to the others by cohesive relations of whatever type.

Chains do not usually occur in isolation, but alongside other chains. However, the mere presence of two or more chains in a text does not guarantee a cohesive effect. Although chains contribute to cohesion in a text, they need to be related to each other somehow. This relationship is called *chain interaction*. The relationships are mostly grammatical, in the Transitivity structure. Hasan establishes a minimum requirement for chain interaction: at least two members of one chain should stand in the same relation to two members of another chain. For a better definition of the interactions, she divides the tokens in a text in three categories:

- Relevant tokens: All tokens that enter into chains, further divided into:
  - Central tokens: relevant tokens that interact
  - Non-central tokens: relevant tokens that do not interact
- Peripheral tokens: Tokens that do not enter into any kind of chain

We are, finally, in a position to define cohesive harmony, which is the function of three phenomena:

- Low proportion of peripheral tokens to the relevant ones
- High proportion of central tokens to non-central ones
- Few breaks in the interaction

Hasan affirms that coherence is a function of cohesive harmony. Our perception that a text is coherent, that it somehow makes sense, is dependent on its cohesive harmony. We will examine the apparently low cohesive harmony of the dialogues, given their low chain interaction.

		English	% English	Spanish	% Spanish
R1	Reference, personal	5	1.08	7	1.13
R2	Reference, demonstrative	108	23.28	49	7.9
R3	Reference, comparative	0	-	3	0.48
S1	Substitution, nominal	1	0.22	0	-
S2	Substitution, verbal	1	0.22	0	-
S3	Substitution, clausal	2	0.43	0	-
E1	Ellipsis, nominal	8	1.72	19	3.06
E2	Ellipsis, verbal	8	1.72	15	2.42
E3	Ellipsis, clausal	4	0.86	5	0.81
L1a	Lexical, same, identical	141	30.39	274	44.19
L1b	Lexical, same, rephrased	67	14.44	43	6.94
L2	Lexical, synonym	23	4.96	33	5.32
L3	Lexical, superordinate	9	1.94	25	4.03
L4	Lexical, subordinate	58	12.5	45	7.26
L5	Lexical, general word	8	1.72	33	5.32
L6	Lexical, collocation	21	4.53	69	11.13
n	Total number of links	464		620	

Table 1. Cohesion types in the corpus.

3. COHESION IN SCHEDULING DIALOGUES

3.1. COHESION TYPES. In the corpus, I identified the types of cohesion described in Table 1. Each instance in the table is to be interpreted as the relationship of one cohesive item to the item that preceded it, that is, as one link. Note that the figures are the overall numbers in the corpus, a total of 60 conversations.

I will first discuss the results for each language, then focusing on a comparison of the two. In English, by far, the resource used most often is lexical cohesion (70.48% of the links are lexical), and more specifically the repetition of the same item. I divided the repetition of same items into two different categories, one where the item is repeated verbatim, and another one where there is some difference in the tie. The reason for this division is that I had the impression that speakers tended to repeat exactly the same terms. They do repeat dates and times quite often. Because the discussion of dates places a high burden on working memory, dates are tossed back and forth and repeated. Dates and times are proposed and abandoned constantly. The speakers need to make sure that they are discussing the same date, and thus they tend to repeat the date at the same time as they mention their availability or unavailability for that particular date.

The same reason justifies the high number in the L4 category, lexical subordinate items. The speakers start out with a two-week window, and then narrow down from

	English	Spanish
Cohesive Links	464	620
Words	6804	9112
Ratio	0.068	0.068

Table 2. Ratios of cohesive links to words in the corpus.

that. Overall, lexical cohesion is the most used device to establish cohesion in these conversations.

After lexical cohesion, reference is preferred. Within reference, English showed no instances of the comparative, only a few of the personal, and a high percentage (23.28%) of the demonstrative type. Instances of personal reference are personal pronouns used to refer to a previous date. The demonstrative can be a modifier or head of a nominal group, sometimes accompanied by a general word, such as *days* or *times*. Another frequent instance of the demonstrative is the use of an adverb, such as *then* or *there*, to refer to the date or place agreed upon.

The last two types of cohesion in order of frequency are substitution and ellipsis. They are both closely related, since ellipsis is substitution by zero. It is clear, from the table, that speakers do not favor these two types. My hypothesis is that they both place a heavy burden on the speakers' minds. It takes extra effort to resolve elliptical references, and, to a certain extent, substitution as well. The figures for the two are very small, making it difficult to draw comparisons. However, ellipsis is used more often. That is, speakers prefer to leave something unsaid than to use a substitute term for it.

Now let us turn to the study of the Spanish data. Most of the phenomena described in English apply to Spanish: higher numbers in lexical cohesion and, within grammatical cohesion, preference for the reference type. There are a few differences worth noting. At first sight, Spanish seems to employ a higher number of ties (620 versus the 464 of the English corpus). However, they are in exactly the same ratio per words as in English (Table 2). Both languages use exactly the same ratio of cohesive devices per word; they only use different proportions of each, as we shall see.

The tendencies in Spanish show, again, a high preference for repetition of the same lexical item, whether in identical form or somewhat rephrased. This type of lexical cohesion is used more frequently than in English, with the identical repetition being the preferred type. The Spanish speakers used lexical cohesion 84.19% of the times a cohesive link was used, a slightly higher percentage than the English one (70.48% for lexical cohesion). In the reference category, we find the most marked difference between the English and the Spanish data. The three types of reference account for 24.35% of the relations in English, whereas it is only 9.52% in Spanish. Demonstrative reference seems to be much more highly preferred in English. In ellipsis the Spanish percentages are closer to the English ones.

The most intriguing category for Spanish was substitution. I could not find any instance of substitution in the Spanish corpus. As I noted above, the frequency of substitution is low in English, but in Spanish it was simply non-existent in my data. Now let us ponder what substitution would look like in Spanish. Consider the following example.

(3) Which book do you want? The red one.
¿Qué libro quieres? El rojo.

In English, the answer uses *one* as a nominal substitute for *book*. In Spanish, the noun is omitted, leaving us with a case of ellipsis. I will not enter into what category *rojo* is, whether adjective or nominalized adjective. But it is clear that there is no substitution. Even if we argue that *rojo* is a substitute for *libro rojo*, we then have a case of repetition, and thus lexical cohesion, but no substitution. Nominal substitution seems to be ruled out for Spanish.

In verbal substitution, English utilizes a number of elements to replace the process or the process plus complements, mainly the auxiliary verb *do* or *do so*, as in (4), where *do* replaces *know her*[3]. In the Spanish counterpart to this example, ellipsis is used instead, leaving the subject pronoun *yo* to stand on its own.

(4) Nobody knows her better than I do.
Nadie la conoce mejor que yo.

This is not to be interpreted as a total absence of substitution in Spanish. The verb *hacer*, a rough equivalent of *do*, is used in some cases. In (5) *hacer* replaces *invadir*.

(5) ¿Van a invadir Kosovo? Parece que eso es lo que van a hacer.
Will they invade Kosovo? It looks like that's what they're going to do.

My impression in this case is that there is a very fine line between this substitution and the lexical cohesion instance where a word is referred to with the class of general words. After all, *hacer* is not as grammaticalized in Spanish as *do* is in English, and we could very well account for (5) by taking it to be an example of lexical cohesion: the use of a general word to refer to a subordinate instance of the same class. Indeed, Casado Velarde (1997) joins substitution and reference under a 'substitution' heading in his description of textual phenomena in Spanish. There he includes lexical proforms—among them the verb *hacer*—but also reference.

Finally, let us look at clausal substitution, where an entire clause is presupposed by an item, in English typically *so* and *not*. In our first example, (6), the English example—adapted from Halliday and Hasan (1976)—substitutes the previous clause by *so*. A possible Spanish translation would use a demonstrative pronoun, *eso* (the rough translation of the Spanish would be 'they say that'). Unless we interpret

	English	Spanish
Immediate	33.45	30.92
Mediated	51.21	56.71
Remote	14.61	12.07
Cataphoric	0.72	0.3
Mediated – Average distance	3.41	4.25
Remote – Average distance	1.16	1.79

Table 3. Average directions and distances (in clauses) per conversation.

the class of demonstrative pronouns as doing double duty, that is, providing reference and substitution, this is an instance of reference.

(6) Is there going to be an earthquake? They say so.
¿Va a haber un terremoto? Eso dicen.

In other instances of short answers, ellipsis again seems to be the alternative in Spanish, rather than substitution. Example (7) is again substitution in English and ellipsis in Spanish. One could argue that *sí* and *no* are the substitute terms for the previous clause. However, the sentences could be completed in Spanish without alterations (*Creo que sí se han marchado; Espero que no se hayan marchado*), which is not possible in English. This leads me to believe that the Spanish speakers are simply leaving the information out.

(7) Did they leave? I think so. / I hope not.
¿Se han marchado? Creo que sí. / Espero que no.

These observations on the nature of cohesion in Spanish are, of course, preliminary. Since my corpus did not contain any instances, I can only speculate about what substitution could be like in Spanish. A larger corpus across different genres and modes would be needed.

Table 3 displays the averages of these types, and the length of the distance in the mediated and remote types per conversation. The table shows that, on average, a conversation will have mediated ties as the most likely type, followed by immediate ones, and then by remote. Cataphoric links are the least common ones. In the mediated type, the average distance between the link under consideration and the first referent—measured in clauses—is 3.41 and 4.25 clauses, for English and Spanish respectively. Thus, even though mediated is preferred over immediate, the length of intervening material in a mediated link is not large: about four clauses. In the case of a remote link, the constraints on distance are more stringent: only one or two clauses on average will separate the two elements in a cohesive tie.

	English	Spanish
Average number of chains per conversation	4.43	4
Average length of chains (number of links)	4.99	6.56

Table 4. Average number and length of chains.

From this we can see that speakers avoid placing an excessive burden on each other's working memories. Elements that are relevant—supposedly the ones in mediated chains—are repeated frequently. Even when there is a remote link, the distances are very close. This might be a characteristic of spoken language, since it is difficult to keep a large number of items in working memory at the same time—seven, plus or minus two, is the magical number, according to Miller (1956). In addition, this particular type of spoken language requires an extra effort from the participants, given the number of dates, times and places that are being passed from one interlocutor to the other one, since chains typically cross over turns.

4. 2. CHAINS: TYPES AND LENGTH. The analysis of cohesive types in the corpus is interesting *per se*. It comes to have full importance only if we consider those links as they are integrated in cohesive chains that run through the text. A cohesive chain is a series of links that are connected to each other through some of the relations already described.

Most of the cohesive chains in the dialogues included days, times and events. In a few instances, they referred to speakers' activities. Numerically, the English corpus contained a total number of 133 chains; the Spanish one 120. The average numbers per conversation are presented in Table 4. Each conversation has, on average, about four chains, with the English ones showing a slightly higher number. However, these chains are longer in Spanish than they are in English.

The most interesting aspect of the chains running through the conversations is that they hardly interact with each other. The typical conversation structure is one where a chain is introduced, discussed and abandoned. After abandoning a chain, a new one is introduced, with little, if any, interaction with the first one. If we follow the chains in most conversations, starting a new one where the previous one finished, the graphical representation of the chains is a right-leaning slope. In order to illustrate this more clearly, I have included a conversation from the English corpus, in (8), which is represented in Figure 1. There are a few minor chains running through the text, which interact to a certain extent with the major ones. The interaction is shown through horizontal arrows in the figure. The main four chains, however, do not interact at all with each other. The links found in the figure are **bolded** in the text.

(8) **s1**: Heather, how's your schedule for, **the fourth**. I'm free all **day**.
s2: /hm/ **The fourth**, I think **that**'s **the date**, **it**'s, **that**'s **today** isn't **it**?
s1: Okay, never mind… What about, /hm/ **the fifth, like at noon**.

(8) **s2**: **The fifth** works for me, I'm free all **day**.
s1: We have two hours between **there**, well, let's see, I have a lab, **at two o'clock**. What about, **Friday the sixth**, I'm free **all day**.
s2: /hm/ **That**'s kind of rough. /um/ Unless we did it like, **eight to ten**. Then I'd have to rush to **a meeting** with Dave. 'Cause **that**'s **at ten**.
s1: Okay. /um/ Let's try **the next week**. On, **the ninth and the tenth**, I can, meet you, from one thirty, on, I'm free. How 'bout you.
s2: /oh/
s1: Heather, I didn't get that last bit **for the ninth** and **the tenth**. I just had blank. Can you repeat **that?**
s2: I'd love to repeat it, 'cause I messed up, anyways, /um/ **the ninth**, I can't, 'cause I have to meet Michael at the airport at three fifteen. But **the tenth**, I just have to prepare **a seminar**. So, what time's best for you on **the tenth**.
s1: /um/ After, we'll see **one fifteen** to the rest of **the day**, I'm free, so, how's your schedule look for **that**.
s2: Well, I can **prepare that seminar all morning**, and, like later on **in the evening**, if need be **so**, how 'bout **two o'clock**.
s1: **That**'s fine, I'll see you **January tenth**, **at two o'clock** for, **two hours**, and I'll see jus, I'll just see you **then**.
s2: Okay, later.

With respect to the type of chain (identity or similarity), most of the chains found in the corpus are of identity, because the same element is repeated and referred to very often. This holds true even despite the low numbers of reference itself: the speakers achieve identity by mere repetition. The identity of the chains depends on the text and the context. The elements hold together much more closely because of the context. The identity of *Tuesday* with any date is only true in the context of a calendar. Again, this is a characteristic of spoken discourse, since spoken discourse is not meant to be repeated or read again—except for, maybe, by the inquisitive linguist.

I will conclude this section with a few remarks on the cohesive harmony of the dialogues. We saw, in Section 2, that cohesive harmony is a description of the cohesion and coherence of text. A text's cohesive harmony is established by high degrees of chain interaction and by ratios of relevant to peripheral tokens. Hasan (1984) also establishes that harmony is maintained if there are no breaks in the chain of interaction. The example above, a typical one, illustrated the very low degrees of interaction in the chains, and how major chains in the text do not interact with each other at all[4]. The scheduling genre seems to be characterized by a process of initiating and abandoning cohesive chains, unlike narrative, for instance, where characters and situations remain constant.

Two explanations are possible at this point. The first one is that spoken language, and particularly scheduling dialogues, have very low or non-existent degrees of

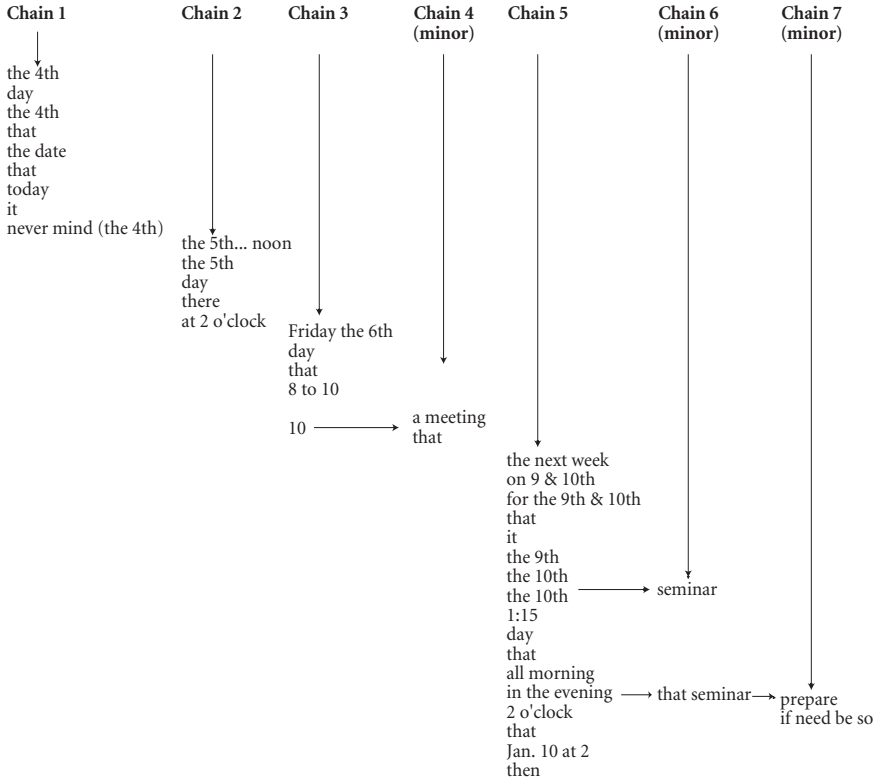

Figure 1. Cohesive chains in Example (8).

cohesive harmony. But these texts are functionally appropriate. The speakers created them for a particular purpose and, given the satisfaction and agreements at the end of the conversation, the texts have fulfilled that purpose. The second explanation points to the need for different measures of cohesive harmony. We might have been relying on measures that apply to narrative and expository texts, whether written or spoken. It could be the case that we need different measures of cohesive harmony for different genres.

4. COHESION AND STAGING. The conversations in this corpus, as instances of a genre, are divided in three main stages: Opening, Task-Performance and Closing. The chains of cohesion, however, run throughout the text. It is my intention in this section to show how the chains behave differently in each of the stages.

Most Openings are short in both languages. The Opening is considered to be over as soon as the speakers start negotiating a date. As a consequence, there are no chain-related elements in an Opening. Example (9) contains the first two turns of

a Spanish dialogue. The Opening is over after the first period, that is, immediately before speaker S1 mentions the need to meet. The first chain begins with the mention of a date (*lunes veinticuatro de mayo*), which is picked up by the other speaker. This chain started two clauses into the Task-Performance stage. There is, then, no chain crossing over from Opening to Task-Performance stage. This is true of all the conversations in the corpus, and for both languages.

(9) **S1**: Bueno Patricia. Parece que tenemos que hacer otra reunión para terminar nuestro asunto. A ver si tiene algunas horas libres el lunes veinticuatro de mayo por la tarde.

S2: El lunes veinticuatro por la tarde va a ser difícil. Porque tengo una reunión...

S1: (Translation) Well Patricia. It looks like we need to have another meeting to finish our business. Let's see if you have some free hours on Monday the twenty-fourth of May in the afternoon.

S2: (Translation) On Monday the twenty-fourth it is going to be difficult because I have a meeting...

As for the Task-Performance phase, all the chains are initiated in it. As we saw in (8), some of them are abandoned for new ones, and some minor ones continue alongside the major ones. The Task-Performance phase is the place where the chains are developed for the most part. This stage is composed of smaller substages, Date or Place Proposals. Since these substages involve the discussion of different dates, it is obvious that we will find a switch over to a new chain as soon as a new date is introduced and discussed.

The chains that originate in the Task-Performance phase do not always cross over into the Closing stage. If they do, they include links that represent the Closing elements in a chain, which include a restatement of the date or a demonstrative reference to it. There are, on average, exactly the same number of links in the Closing section in the Spanish and in the English dialogues: 2.33 links. These links *always* belong in the last major chain of the dialogue, and are most often instances of repetition (lexical cohesion) or demonstrative reference. Such is the case in (10), which contains the last two turns of a conversation. The Closing begins immediately after speaker S1 has accepted the date (*couldn't be better*). The last SDU in her turn repeats the date and the time that have been discussed previously in the chain (*eleven to one on the twenty-sixth*). Speaker S2 responds by referring to that time (*then*), an instance of demonstrative reference.

(10) **S1**: Couldn't be better. And then you know maybe if we if we're, over a little bit, you know, it would be okay. So, eleven to one, on the twenty-sixth.
S2: Sounds great. See you then.

The most interesting phenomenon regarding cohesion and staging is the presence and evolution of cohesive chains within the Task-Performance stage. Within the Task-Performance stage proposals for dates to meet are tossed back and forth until one of them suits both speakers. The different cohesive chains that run through the Task-Performance stage are closely related to the beginnings and ends of the Proposal stages. That is, a new Proposal initiates a chain, and the rejection or completion of the Proposal stage terminates that chain. If we look again at the chains in Figure 1, we observe that the major chains running through the text correspond to different Proposals. Minor chains do not correspond closely with Proposal stages, because they are supporting, or in close relation to, a major chain. Again, cohesive harmony is violated: there are whole parts of the text that do not interact at all with each other. Rather, non-interacting chains mark the transition from one stage to the next one.

The conclusion here is that we can predict a new stage given a new chain, and vice versa. We could use this knowledge in different applications, such as automatic recognition of stages for information retrieval. If we know that the participants in a conversation began a new cohesive chain, we can predict that they are in a different generic stage of that conversation. In the opposite direction, knowing that we have started a new stage usually means that anaphora resolution will take place within that stage, because references will usually be only to elements in the same cohesive chain.

5. CONCLUSIONS. This paper has presented a study of cohesion in a bilingual corpus. Cohesion is interpreted according to the Halliday and Hasan (1976) model. Some adaptation of that theory was necessary in the analysis of Spanish conversations. The most notable application revealed the questionable presence of substitution in Spanish.

The comparison of the number of cohesive links used for each language discovered exactly the same ratio (0.068) of cohesive elements to words in both English and Spanish. The types of cohesive elements that the speakers avail themselves of were also very similar in both languages. Lexical cohesion, and specifically the repetition of the same item, is the most widely used type. Lexical cohesion was followed by the use of reference.

In terms of direction of reference, both languages preferred anaphoric reference, with cataphoric reference occurring a very small percentage of the time. The anaphoras were most frequently mediated, that is, they entered into a relatively lengthy chain of elements. However, the links are often activated, making the distance between a link and its mediating element only slightly longer than four

clauses. After mediated reference, in both languages, comes immediate reference. All of this shows that the speakers avoid establishing links between elements located far from each other. I argue that this is a characteristic of spoken language in general, and of this genre in particular.

Finally, and with regard to chains, there are also striking similarities between the two languages. Each conversation contains about four chains (4.43 in English and 4.00 in Spanish). Those chains rarely interact with each other, because one is abandoned as soon as a new one is initiated. This holds true for major chains involving times and dates to meet. There are also minor chains throughout the dialogues, which show slightly higher degrees of interaction. Although the Spanish dialogues contain lower numbers of chains than the English ones, those chains are lengthier (an average of 4.99 links in English versus 6.56 in Spanish).

I conclude with some remarks on the presence of cohesion according to the different stages. The Opening phases of the conversations did not establish cohesive links with the rest of the conversation, which contained the most number of links and chains. The Closings only contained an average of 2.33 links to the last major chain running through the dialogue, the same average for both languages. The Proposal stages, repeated within the Task-Performance stage, contain new cohesive chains, usually unrelated to other major chains. As a result, we can say that the presence of a new chain is a good index of a new Proposal stage.

---

[1] Thanks to the Interactive Systems Laboratory and to Alex Waibel, its director, for permission to use the corpus.

[2] The transcripts include a number of conventions introduced by the transcriber, to reflect every sound produced during the conversation. For ease of reading, I have deleted most of them, only leaving some that seemed relevant: */hm/* and */um/* indicate a hesitation on the part of the speaker. Stretches of talk accompanied by laughter are surrounded by */begin_laugh/* and */end_laugh/*. *pause* indicates a period of time from 0.5 to 2 seconds with no sounds at all. Turns are simply marked as 's1' and 's2' to indicate 'Speaker 1' and 'Speaker 2'.

In addition, transcriber comments include intonation, marked with either a comma (,), a period (.) or a question mark (?) at the end of the corresponding section of speech. These markings do not reflect, nor are influenced by, sentence structure. The speaker may have the intonation of a statement whether he or she is, in fact, asking a question. He or she may have the falling intonation typical of the end of a sentence (reflected in a period) after a collection of words that do not, in any way, resemble a grammatically correct or complete sentence.

[3] Note that, in English, there is an ellipsis alternative to the substitution, further complicated with the shift of case, from *I* to *me*: 'Nobody knows her better than I/me'.

[4] Cohesive harmony has as prerrequisite the interaction of chains. Once interaction is established, a measure of central, non-central and peripheral tokens establishes the harmony level. Given the low interaction, I did not carry out the full analysis.

REFERENCES

Casado Velarde, Manuel. 1997. *Introducción a la gramática del texto del Español*. Madrid: Arco Libros.

Halliday, M.A.K. & Ruqaiya Hasan. 1976. Cohesion in English. London: Longman.

———. 1985 *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Hasan, Ruqaiya. 1984. Coherence and cohesive harmony. In *Understanding reading comprehension*, ed. by James Flood, 181–219. Newark, de: International Reading Association.

Mann, William C. & Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3):243–81.

Martin, J. R. 1985. Process and text: Two aspects of human semiosis. In *Systemic perspectives on discourse*, Vol. 1, ed. by James Benson & William Greaves, 248–74 Norwood, nj: Ablex.

Miller, George A. 1956. The magical number seven, plus or minus two. *Psychological review* 63:81–97.

Parsons, Gerald. 1996. The development of the concept of cohesive harmony. In *Meaning and form: Systemic functional interpretations (Meaning and choice in language: Studies for Michael Halliday)*, ed. by Margaret Berry, Christopher Butler, Robin Fawcett & Guowen Huang, 585–99. Norwood, nj: Ablex.

Taboada, Maite. forthcoming. Rhetorical relations in dialogue: A contrastive study. In *Discourse across languages and cultures*, ed. by C. L. Moder & A. Martinovic-Zic. Amsterdam: John Benjamins.