Clause-linkage techniques as a symptom of discourse traditions: methodological issues and evidence from Romance languages

Johannes Kabatek, Philipp Obrist, and Valentina Vincis

Eberhard-Karls-Universität Tübingen

http://www.sfb441.uni-tuebingen.de/b14/

Abstract:
Throughout the last two decades, corpus linguistics’ methods have continuously been used to account either for the general historical evolution of particular linguistic phenomena or for their variation across genres, but rarely for both at the same time. In the present contribution, we will first introduce the notion of discourse traditions, a theoretical concept introduced by scholars in the field of Romance linguistics for the description of textual traditions. The main postulate associated with this concept is that textual traditions must be regarded independently from the diachronic evolution of language as a system; nonetheless, they may have a strong influence on it.
In the second step, clause linkage techniques, commonly known to be highly symptomatic for the identification of textual traditions, are analyzed. A quantitative approach will be used on a series of historical Romance corpora: juridical texts from 13th century Spain, an Old Spanish Bible corpus, a corpus containing texts of various Old Spanish genres and, finally, a Rheto-Romance newspaper corpus from the last two centuries. Results show the importance of genre differentiation in corpora in two different dimensions: “vertically,” distinguishing texts traditions with globally different characteristics; and, “horizontally,” identifying differences within the single texts.

1. Introduction

In the last years, the idea that languages are synchronically homogeneous has been criticised in several studies that have claimed that there is not only linguistic variation due to different varieties of a language but also due to different textual traditions. Not all linguistic elements appear in the same way in all texts, and there are elements that tend to be attached to certain textual genres but not to others. This is of course true for lexical items, but it seems to be valid as well, at least to some degree, for grammatical elements. A well-known example from the Romance languages is the French passé simple, a verbal tense almost entirely restricted to a series of written text types in France. From a diachronic view, this means that innovations can emerge in some genres without affecting others at first, but also that a loss of grammatical forms does not have to affect all genres at the same time (see Kabatek 2005a;
Koch 2008; Pons 2008). Another example of this phenomenon can be seen in the Spanish subjunctive future form *amare*, which was still in general use in Old Spanish, but had almost completely disappeared by the beginning of the eighteenth century; however, it has been preserved until the present, almost solely due to its use in juridical texts.

In the last years, several scholars in Romance linguistics have introduced and developed the concept of *discourse traditions*, by taking the textual variation in a given language into account (Koch 1997; Oesterreicher 1997; Wilhelm 2001; Aschenberg 2003; Kabatek 2005b; Kabatek ed. 2008). Discourse traditions are textual traditions that are related not only to certain elements of content, but also to the lexical and grammatical items predominantly used in that tradition. The notion of discourse traditions was introduced as a purely theoretical concept with convincing arguments, intuitively applied to a series of textual traditions, but without a properly elaborated methodology. Several examples, generally based on detailed philological analysis, have shown that textual traditions could strongly determine the choice of lexical and grammatical elements in a text. However, the problem of detailed philological analyses is that they cannot be applied to larger text corpora. In order to differentiate and to compare discourse traditions we need systematic criteria like those applied by scholars such as Douglas Biber and others for the identification of text genres both in synchronic and diachronic corpora (see Biber 1993, 2004; Biber, Conrad, and Reppen 1998; Biber, Davies, Jones, and Tracy-Ventura 2006; Parodi 2007, among others).

Operating in a fashion that is similar to Biber’s method, but with a series of crucial differences, the present paper will present some results of genre-differentiating corpus analyses that are not based—as in Biber’s work—on a large number of different elements that are analysed statistically in order to show which of the elements can be regarded as relevant for genre differentiation. Instead, our method is limited to the analyses of a reduced number of elements considered to be particularly symptomatic of genre differentiation. The basic
hypothesis is that elements that fulfill the task of linking clauses can strongly indicate certain particularities of texts and textual traditions. Evidence of this phenomenon becomes particularly clear when one pays attention not only to the elements themselves, but also to their relative frequency in the texts. The following pages will present some of the results of a research project on the relationship between clause-linkage techniques and discourse traditions, presenting empirical data from diachronic corpora of two Romance languages along with some general implications for theory and methods.

The section following this introduction is dedicated to the presentation of the framework of discourse traditions theory and in particular to analyses of clause linkage (or “junction,” in Raible’s terminology, see Raible 1992: 28) as applied in the Tübingen project B 14/SFB 441.1 Section 3 offers the first results of textual differentiation in Old Spanish texts, in which texts that belong to the same genre (legal texts) show strong internal differences due to three different historical backgrounds, representing thus three different discourse traditions within the same genre. In section 4, a larger Old Spanish corpus will be presented and the question of the relationship between diachrony and textual tradition will be discussed. Section 5 is dedicated to the internal differences in a single text. In this section, different versions of Old Spanish Bibles will be compared in order to show, on the one hand, that clause-linking elements can be attributed to textual structure, and on the other hand, that different textual models (in this case, Latin and Hebrew origins) can lead to rather different results in each respective Bible version. Finally, section 6 will show how clause-linkage techniques have evolved in Rheto-Romance newspapers from the nineteenth century to the present, and how a rather stable textual tradition changes over time from a more subordinated to a more nominalised style. The general objective will be to show: 1) the relevance of genre

1 The project B 14 Discourse traditions of Romance languages and multidimensional analysis of diachronic corpora, directed by Johannes Kabatek and with collaboration of Philipp Obrist and Valentina Vincis, is part of the linguistic collaborative research centre SFB 441 Linguistic Data Structures: On the Relation between Data and Theory in Linguistics at the University of Tübingen. The research centre is funded by the German Research Council (DFG). See http://www.sfb441.uni-tuebingen.de/index-engf.html
distinction and the notion of discourse traditions in relation to synchronic and diachronic studies; and 2) the possibilities for and limitations of the analysis of clause-linkage patterns for textual differentiation.

2. On discourse traditions and the qualitative-quantitative method

The theoretical framework for the empirical and methodological issues discussed in this paper is discourse tradition theory. The notion of discourse traditions was first presented by Peter Koch and Wulf Oesterreicher (1997), and has earlier roots in Eugenio Coseriu’s functionalism (Coseriu 1988: 161–181) and in Schlieben-Lange’s (1983) theory of historical pragmatics. During the last decade, a series of papers and monographs have further elaborated the basic notion of discourse tradition in its theoretical and methodological aspects and have demonstrated in several empirical contributions its useful application, above all, in the field of historical linguistics. However, since the basic assumption of the framework is that every utterance might be determined not only by grammar and lexicon but also by textual tradition, discourse tradition theory is by no means limited to diachronic issues but fundamental for synchronic linguistics as well.2

The basic assumption of discourse tradition is very simple and evident. Producing utterances is not simply an activity that results from a certain grammar of a language, but also (and simultaneously) according to concrete utterances that have already been produced: utterances that are part of the tradition of a community, or discourse traditions. What seems so simple and evident has a number of important consequences, of which we will mention only a few:

2 The insistence on text tradition is not new at all; see e.g. Hockett’s claim that tradition is a linguistic universal: “3.5 Every human language has tradition (…). If we design and build a collection of machines that communicate among themselves with a language, this property will be lacking.” (Hockett 1966: 12).
- The tradition of the texts does not only lead to the *addition* of elements (such as genre-specific markers, textual identifiers, or textual formulae etc.) but also determines the *selection* of elements in texts, and not only lexical items but also grammatical elements.

- Synchronic variation in texts is not only a matter of differences between dialects or varieties, but also of differences between textual traditions.

- Even if the origin of a textual tradition might often be described by means of concrete pragmatic needs, discourse traditions are more than only situation-specific coincidences: it might be true that there are speech acts that are strongly determined by situational constellations (asking for a light for a cigarette, for example), but the concrete linguistic form (“do you have a light?”) of resolving the social task is very frequently taken over by formerly pronounced texts and is not spontaneously invented.

- Discourse tradition is a general notion for tradition in language, and it includes all linguistic results of tradition, from simple formulae up to complex literary traditions.

- Synchronic and diachronic variation must be differentiated according to different discourse traditions. In diachrony, by looking more closely at discourse traditions, it seems possible to shed light on abstract views that regard linguistic change as putative invisible-hand phenomena such as the s-curve or other apparently regular evolutions (Kroch 2001).

- Discourse tradition is not only a relevant notion for the repetition of texts, text types or genres, formulae, and so on, but for all linguistic phenomena related to traditions of utterances, including a significative absence of textual repetition in pragmatically analogous situations, social stratification and group phenomena correlating with textual traditions, politeness, textual interferences, and so on.

From its first presentation onwards, the notion of discourse tradition has appeared in theoretical as well as in empirical studies. Koch and Oesterreicher presented it as a fundamental contribution to linguistic theory, and the ensuing elaborations of the concept
(see, among others, Wilhelm 2001; Aschenberg 2003; Kabatek 2005a; López Serena 2007; Pons 2007; Sáez Rivera 2007) have elaborated and added a series of aspects. At the same time, several empirical studies have identified discourse traditions in different Romance languages (Wilhelm 1996; Stoll 1997; Oesterreicher, Stoll, and Wesch 1998; Wilhelm 2001; Schmidt-Riese 2003; Kabatek 2005b; Ciapuscio, Jungbluth, Kaiser, and Lopes 2006; Guzmán Riverón 2007; Castilho in press) and others have shown the importance of this notion for historical syntax (Simões 2007; Kabatek 2008). The elements that have been identified as relevant for discourse traditions are, on the one hand, peripheral elements such as explicit notions of a certain textual tradition, formula or the overall textual form, and on the other hand, certain lexical or grammatical items.

The project we are currently presenting has its starting point in a critical view on monolithic diachronic studies that consider diachrony as a linear evolution of one homogeneous object, with variation other than the diachronic one being nothing but a matter of noise in the data that can be compensated for by large quantities of data. At the same time, we point out that the detailed philological analyses of individual texts, perfect as they might be, cannot serve as a basis for comparison with other texts: to be able to compare and to distinguish by means of a tertium comparationis, abstraction is necessary at least to some degree. The first step towards this abstraction consists of the choice of possibly relevant elements from textual traditions and their comparison in different texts. In many cases, a discourse tradition can be identified simply by certain keywords contained in a text. In other cases, however, it is not only the mere appearance of a word but also the frequency of its appearance that helps us to identify the discourse tradition. The twenty most frequent words of a text may give strong evidence for its discourse tradition. Our claim is that the same holds for certain grammatical elements and that their relative frequency in texts should correlate with different discourse traditions. Changing the point of view from reception (and
observation) to production, this means that speakers will make their choices of grammatical
variants according to the textual tradition they wish to represent. The relative frequency and
the distribution of elements along a text will be considered as grammatical indicators of
discourse traditions, accompanied by all the other indicators (lexical, formulaic, and so on).

According to the findings on corpora and genres applied to different languages by
Biber and others, the continuum between orality and literacy is the most salient factor that
determines the choice of elements in different texts. Koch and Oesterreicher stated in their
seminal work on the continuum between what they call “language of immediacy” and
“language of distance” (Koch and Oesterreicher 1985: 21; see also Koch and Oesterreicher
1990, 1994) that with regards to clause-linking, there is a strong correlation between this
continuum and a scale between “aggregation” and “integration.” Whereas at the pole of
aggregation (correlated with “language of immediacy”), we can find simple juxtaposition of
clauses, at the pole of integration (correlated with “language of distance”) we will find clause
integration by techniques of condensation, like e.g. in nominalising constructions. In 1992,
Wolfgang Raible, in the tradition of the Cologne school of typology, offered a complex
type theory of what he calls “Junktion” (Raible 1992: 28; i.e. clause linkage, see Raible 2001). He
suggests the following stages between aggregation and integration (Raible 2001: 595), some
of which are illustrated by English examples:3

<table>
<thead>
<tr>
<th>Stage</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>simple juxtaposition</td>
<td>Joan remains at home. She is ill.</td>
</tr>
<tr>
<td>II</td>
<td>junction by phrasal adverbs</td>
<td>Joan is ill. This is why she remains at home.</td>
</tr>
<tr>
<td>III</td>
<td>coordinating conjunctions</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>subordinating conjunctions</td>
<td>Since Joan is ill, she remains at home.</td>
</tr>
<tr>
<td>V</td>
<td>infinite verbs (gerunds or infinitives)</td>
<td>Being ill, Joan remains at home.</td>
</tr>
<tr>
<td>VI</td>
<td>prepositional groups,</td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>Nominalisations</td>
<td>On account of her illness, Joan remained at home.</td>
</tr>
</tbody>
</table>

Raible combines this continuum between aggregation and integration with a second
axis of semantic categories. This second axis, based on a cognitive and logical scheme of

3 The original examples of the whole classification were French. In Raible 2001, the above English examples are
given for some of the levels. Not all levels must have a representation in all languages; Raible considers his
continuum to be universal, with concrete forms differing from one language to another.
increasing relational complexity, arranges the semantic relationship between clauses on an ordered scale: condition is considered to be less complex than cause; “counter-cause” (the concessive relationship) is considered to be more complex than cause, and so on (Raible 1992: 131.; see fig. 1). This cognitive scheme (see Kortmann 1997, 2001) is related by Raible to language acquisition, language change, and the difference between orality and literacy.

Figure 1: Two dimensions of “junction,” according to Raible (1992)

In an onomasiological perspective, the junction scheme is a field of two universal continua; in a semasiological perspective, every language offers a series of elements (junctors) and techniques of clause linkage that can be arranged in accordance with the two dimensions. The junction scheme can be applied to a number of linguistic fields: typologically, language comparison can be based on the junction techniques languages offer; in a diachronic perspective, evolutions of junction techniques and junctors can be observed. The complete inventory of junctors in a language can be described as a characteristic pattern. In synchrony, language patterns and textual patterns can be distinguished, the former consisting of the complete clause-linking grammar of a language, the latter consisting of the clause-linking elements found in a text. These textual junction patterns will be considered, in our own
studies, as relevant for the textual structure (see section 4), and, as we will see in the next section, for different discourse traditions.

3. The first step: Discourse traditions in Old Spanish texts

In a detailed linguistic analysis of Old Occitan and Old Spanish juridical texts (Kabatek 2005b), we found striking textual and linguistic differences between several Spanish legal texts from the thirteenth and fourteenth centuries. Even if all of these texts could be categorised as, or even had categorised themselves as, laws, there are still fundamental differences that could not be explained in terms of diachronic evolution: even within a very clearly defined genre, there were important subcategories. If we consider the fact that many historical studies suppose that a simplistic genre distinction between general text types like juridical prose, literary prose, poetry, etc. is enough to give a representative view on textual variation in a language, the existence of subcategories seems to be an important observation. The explanation we offered for the variation was a historical one: in thirteenth-century Castile, three different law systems co-existed, the first one being based on descriptions of cases (the so-called fazañas) with implicit laws, the second one offering general norms or more abstract laws (fueros), and the third one deriving from the Renaissance of Roman law from the twelfth century onwards (for the “Renaissance of the 12th Century”, see already Haskins 1927) and based on a complex juridical system with elaborate institutional and meta-juridic instructions. The texts, and accordingly, the linguistic elements that represent the three traditions, are dissimilar: whereas the case laws are based on simple aggregative case descriptions with short, unconnected simple sentences, the abstract Spanish local laws tend to be lists of subordinated, conditional constructions. The texts based on Roman law, finally, are elaborated and complex texts with more integrative structures and categories that are more semantically complex than those found in the fazañas and the fueros. Applying the Raible
scheme to the clause-linking elements in our three juridic traditions, we found strong differences in the junction patterns they presented. The differences were even clearer when measuring the *relative frequency* of these elements, since even in the least complex tradition of case law, some isolated highly integrative exceptions (probably due to Latin influence) appeared. The idea of measuring the relative frequency was adopted from Biber (1993). As in Biber’s studies, we calculated the relative frequency per 1000 words. The following two schemes offer some general results of different junction patterns in three Old Spanish juridic traditions (see Kabatek 2005a, 2005b).

Figure 2: Relative number (per 1000 words) of clause-linking elements in three Old Spanish legal texts ordered by semantic relationships between clauses. The numbers 1-9 represent degrees of increasing semantic complexity (1=addition to 9=concession).

![Figure 2](image)

Figure 3: Relative number (per 1000 words) of clause-linking elements in three Old Spanish legal texts ordered by degrees of *integration* (syntactic density) between clauses. The numbers from II-VII represent degrees of increasing integration (II=adverbs; III=coordination; IV=subordination; V=infinite verbs; VI=PrepP with infinite verbs; VII=nominalizations).

![Figure 3](image)
The result of this study was that the junction patterns differed clearly in the three texts and that these differences were not purely idiosyncratic ones: the three texts can be considered representative of the three legal traditions we identified. The comparison of the junction patterns of our three texts to those of other Old Spanish legal texts confirmed that the patterns could be regarded as representative features of the three traditions, since similar patterns were found in other texts belonging to each of them. This means that there existed not only three legal traditions in thirteenth-century Spain, but that they corresponded to three discourse traditions with respective linguistic characteristics which we were able to identify. The questions arising from this first study were the following: 1) Could the observations made in the case of the legal texts be generalised?; 2) To what degree is junction-pattern differentiation a helpful instrument for the identification of discourse traditions?; and 3) Can we distinguish diachronic variation from textual variation? These and other questions will be presented in the following sections.

4. Clause linkage, diachrony, and discourse traditions in Old Spanish

In order to see how clearly junction patterns can be used to identify different textual traditions, we expanded our perspective on Old Spanish by enlarging the corpus in two directions: other genres were included in the analysis on a synchronic axis (the thirteenth century); and, texts from the early thirteenth until the late fifteenth century were added on a diachronic axis. As to the methodology, a computational tool, the program TraDisc, was elaborated in order to make larger portions of texts accessible. TraDisc is a semi-automatic annotating instrument that allows for the creation of textual junction patterns. This program must be trained by introducing the junction techniques of a language. After that, the ambiguous cases must be checked manually in order to extract a textual junction pattern.4

---

4 Program and help files available at: http://www.sfb441.uni-tuebingen.de/b14/tools.html
The set of Old Spanish texts we chose for a first enlargement of the corpus is as follows:

- The *Cantar de Mio Cid* (CMC), a medieval lay on a hero of the Reconquista, supposedly written down at the beginning of the thirteenth century.
- The *Milagros de nuestra Señora* (Miracles of Our Lady, Mil) by Gonzalo de Berceo, a devotional text in verse from the middle of the thirteenth century (between 1246 and 1254), attached to the literary genre of the *Mester de Clerecía* (Ministry of Clergy).
- The *Lapidario* (Book of Stones, Lap), from 1250, and the *Libro de los judizios de las estrellas* (Book of Judgments of Astrology, Est), from 1254; these are two of the scientific prose works attributed to Alfonso X, both mainly translated from earlier Arabic works, and partially updated.
- The *Conde Lucanor* (Count Lucanor, CL), a prose work with didactic and moralistic purposes, written by don Juan Manuel between 1325 and 1335.
- The *Libro de Buen Amor* (Book of Good Love, LBA) by Juan Ruiz, the Archpriest of Hita, contemporary to the *CL* (from between 1330 and 1343); another example of the *Mester de Clerecía*, mainly in verse.
- The *Celestina* (Cel) from 1499, a novel that is seen as marking the end of medieval literature and the beginning of the literary Renaissance in Spain.5

We will now turn to an analysis of the data. Again, the junctors have been calculated per 1000 words (with a portion of 10,000 to 30,000 tokens per text being analysed) and they have been ordered according to a modified scheme based on the two Raible dimensions:

---

5 The extent to which these texts are representative may be questionable. As the project advances, we are hoping to include more and more texts; however, two problems with such inclusion are the limited number of text series in different traditions in this early period of the Spanish language and the lack of free access to reliable electronic full texts, access which is necessary for annotation. Leaving corpus design issues aside, some first observations seem to be nonetheless possible.
At first glance, the data seem not to allow for very clear conclusions. However, a closer look allows for the following very general observations about the results:

- There is a diachronic shift to higher scores of more integrative junction (subordination, infinite constructions, and nominalisations). While three out of four texts from the thirteenth century range show a ratio of less than 26 per 1000 higher density clause-linkage elements, the three later texts show a ratio of 37 and above.

- Variation within genres is much higher than originally expected: Mil and LBA may be partly considered to pertain more or less to a similar genre (Mester de Clericia); and Lap and Est are usually taken as examples of scientific prose. Both pairs show significant differences, and while the former pair is separated by a span of 80 years, Lap and Est are coetaneous.

### Table 1: Junction analyses in a series of Old Spanish texts

<table>
<thead>
<tr>
<th></th>
<th>CMC</th>
<th>Mil</th>
<th>Lap</th>
<th>Est</th>
<th>CL</th>
<th>LBA</th>
<th>Cel</th>
</tr>
</thead>
<tbody>
<tr>
<td>~1200</td>
<td>~1250</td>
<td>~1250</td>
<td>1254</td>
<td>~1330</td>
<td>~1330</td>
<td>1499</td>
<td></td>
</tr>
<tr>
<td>epic</td>
<td>hagiographic</td>
<td>scientific</td>
<td>scientific</td>
<td>didactic</td>
<td>hagiographic</td>
<td>fictional</td>
<td></td>
</tr>
<tr>
<td>Coordination</td>
<td>22.9</td>
<td>18.3</td>
<td>70.6</td>
<td>60.2</td>
<td>63.8</td>
<td>24.3</td>
<td>26.3</td>
</tr>
<tr>
<td>II-III</td>
<td>54.5%</td>
<td>46.6%</td>
<td>65.4%</td>
<td>70.2%</td>
<td>61.5%</td>
<td>39.3%</td>
<td>38.7%</td>
</tr>
<tr>
<td>Subordination</td>
<td>11.4</td>
<td>13.0</td>
<td>26.3</td>
<td>24.8</td>
<td>26.0</td>
<td>17.3</td>
<td>19.3</td>
</tr>
<tr>
<td>IV</td>
<td>27.1%</td>
<td>33.1%</td>
<td>24.4%</td>
<td>28.9%</td>
<td>25.1%</td>
<td>28.0%</td>
<td>28.4%</td>
</tr>
<tr>
<td>Gerunds</td>
<td>3.4</td>
<td>2.1</td>
<td>5.5</td>
<td>0.5</td>
<td>3.9</td>
<td>3.0</td>
<td>9.7</td>
</tr>
<tr>
<td>V</td>
<td>8.1%</td>
<td>5.3%</td>
<td>5.1%</td>
<td>0.6%</td>
<td>3.8%</td>
<td>4.9%</td>
<td>14.3%</td>
</tr>
<tr>
<td>PrepP</td>
<td>4.3</td>
<td>5.9</td>
<td>5.6</td>
<td>0.2</td>
<td>10.0</td>
<td>17.2</td>
<td>12.7</td>
</tr>
<tr>
<td>VI-VII</td>
<td>10.2%</td>
<td>15.0%</td>
<td>5.2%</td>
<td>0.2%</td>
<td>9.6%</td>
<td>17.8%</td>
<td>18.7%</td>
</tr>
<tr>
<td>Total II-VII</td>
<td>42.0</td>
<td>39.3</td>
<td>108.0</td>
<td>85.7</td>
<td>103.7</td>
<td>61.8</td>
<td>68.0</td>
</tr>
<tr>
<td>IV-VII</td>
<td>19.1</td>
<td>21.0</td>
<td>37.4</td>
<td>25.5</td>
<td>39.9</td>
<td>37.5</td>
<td>41.7</td>
</tr>
</tbody>
</table>

At first glance, the data seem not to allow for very clear conclusions. However, a closer look allows for the following very general observations about the results:

- There is a diachronic shift to higher scores of more integrative junction (subordination, infinite constructions, and nominalisations). While three out of four texts from the thirteenth century range show a ratio of less than 26 per 1000 higher density clause-linkage elements, the three later texts show a ratio of 37 and above.

- Variation within genres is much higher than originally expected: Mil and LBA may be partly considered to pertain more or less to a similar genre (Mester de Clericia); and Lap and Est are usually taken as examples of scientific prose. Both pairs show significant differences, and while the former pair is separated by a span of 80 years, Lap and Est are coetaneous.
Finally, there seems to be no direct relation of concurrence between the more aggregative and more integrative junctors in Old Spanish. The comparison of *Lap* and *Est* rather suggests that clause linkage by different means works fairly independently, since both texts show high frequencies of aggregative junctors, but considerable divergence of hypotactic ones. The first finding, a growth in the more integrative sector, is the finding that fits best with the traditional view on the emergence of written Romance languages and with the further elaboration, or *Ausbau*, of these languages—the idea that these languages are written traditions that were and are in the process of becoming more and more diversified and elaborate. But these data—and a closer look at the texts—also suggest that an overall comparison of clause-linking patterns in texts only makes sense if we are dealing with rather homogeneous ones, as was the case with the legal texts presented above in section 3. From the second half of the thirteenth-century onwards, more heterogeneous texts appear, combining and elaborating former traditions. Thus, the application of our analysis to more heterogeneous texts makes it necessary to differentiate the *progression* of clause-linking elements. In some of our texts, considerable differences could be found between separable parts of the texts. This is why easily identifiable sub-texts were regarded as more or less representative of independent discourse traditions. An interesting example is the tradition of prologues of Old Spanish texts: because of the rhetorical necessity of explaining the *intentio* and the *utilitas* of the text in the foreword, the clause-linking elements found frequently in this context express, on the semantic axis, causality and finality (Estrada García 2007).

In some texts, like the fourteenth century *El Conde Lucanor*, a collection of *exempla*, the rather stereotypical scheme of introduction, exemplum, and moral conclusion, implies a high degree of diversity in the corresponding clause-linking elements. This is why Kamilla Tarasiewicz (2006) proposed, as a complementary element to the global junction pattern
analysis, depicting the junction progression in a *junctogram*, a horizontal graph that follows the tokens of a text to show the position of the different clause linkers within the text. The junctogram, which has been integrated as an export function into the program TraDisc, allows us to see where the linking structure of the text changes considerably; it is possible, for example, to distinguish more narrative from more argumentative passages within texts with the help of this tool.

The question that arises from the junctogram visualisation is whether there is a methodological possibility of distinguishing clause-linking elements that derive from the argumentation structure of a text from those serving pragmatic needs (thus somewhat independent from other factors), or if we can show that other factors can play a role. In short: does textual difference derive only from the different content in the texts or can the same content be expressed in different ways according to different textual traditions or styles? This question will be discussed in the following section.

5. Clause linkage and text structure: The case of Old Spanish Bibles

A fortunate circumstance made it possible that in 2007 the Tübingen project on clause linkage and discourse traditions could establish research collaboration with the Spanish Bible Corpus project coordinated by Andrés Enrique-Arias at the University of the Balearic Islands (Spain). By merging the two approaches, it was possible to apply junction analysis to large parallel corpora of different Old Spanish Bible versions. When comparing different versions of the same text, very similar junction patterns and junction progressions should be expected. In fact, as can be shown in Figure 4, the clause-linking elements are almost identical in four different Old Spanish Bible translations of an example extracted from the book of Numbers (22:21-35):

---

6 This joint project is funded by the DAAD/MEC program PPP/Acciones Integradas.
What we can see in this example is a rather homogeneous textual progression in all versions: whereas in verses 21-28 and 31, narrative passages present, above all, additive elements of coordination (III-1) and other elements appear only sporadically, in the two passages marked by the two black rectangles – a passage corresponding to the dialogue between Balaam and the donkey, in contrast to the surrounding narrative context – more integrative structures (degree IV, 3, 5, 6, 7, 8) can be found in all versions.

However, there are also passages where not all versions proceed in the same way. In the following example, one of the versions offers a different solution than the rest:
Here we can see in the passages highlighted in dark grey that one of the versions, E8, shows slightly different solutions in several cases (in 12 cases in this example). In the light-grey passages, all the versions show the same solution or, as in 4 cases, one of the other versions has a divergent solution.

Even a “blind” analysis of the different Bible versions based only on objective visualizations of junctors shows that version E8 seems to be different from the rest, and not only in the example presented here, but almost systematically throughout the text. What we can see here is confirmed by the historical information on the exceptionality of E8: as opposed to the other versions that were translated from Hebrew, E8 is a thirteenth-century version translated from the Latin Vulgata. The result is perfect Old Spanish in both cases, but the way of saying things is slightly different according to the models behind the texts.

The junctogram analysis of nine Old Spanish Bible versions also allowed for a global comparison of the similarity between them.¹ The objective abstract projection of junction patterns in this case confirms the results of detailed philological analysis: the more frequent

¹ For a comparison of similarity between Old Spanish Bibles see also Enrique-Arias (2006).
that similar junction patterns between the different versions are, the more these versions have
to do with each other historically. E3 and Aj, which are almost 100% identical, are two
manuscripts of the same version; GE1 and Faz show disparate values since they are not only
translations but also elaborations of the original text, with a number of additions and
comments in the case of GE1 and with considerable reductions in Faz. The similarity
between GE1 and E8 might be explained by the fact that both texts have the Latin Vulgata as
model for the translation.

![Similarity between different Old Spanish Bibles according to clause-linkage techniques (overall score)](image)

Figure 6: Similarity between different Old Spanish Bibles according to clause-linkage techniques
(overall score)

The comparison of several versions of the same text (different versions of the Old Spanish
Bible) has shown some interesting results of clause-linkage analysis when completed with the
more detailed junctogram: the junctogram makes it possible to identify junction patterns that
are based on the textual structure of content and it can show deviations from the rule when
several versions of one text are compared. In the next section, a different stability, that of a textual genre, will be taken as a starting point for a diachronic study of junction patterns.

6. Clause linkage in Rheto-Romance newspapers

Throughout the previous sections, we have changed our focus from a general overview on junctors in the Old Spanish period to a more precise direction: variation and change within a determined passage of the Bible re-elaborated in different traditions. Notwithstanding the fact that we know much about the literary origins of the Old Spanish texts or the Bible translations, it remains difficult to reconstruct the exact stylistic effect that the different junction techniques might have created.

For this reason, the last section of this paper will be dedicated to a more modern genre. We will, nonetheless, evaluate it in its historical context: the evolution of journalistic style in Rheto-Romance, and more precisely, Romontsch sursilvan, one of its regional variants spoken in the Upper Rhine valley in the Swiss canton of Grisons. While this variant has a literary history of less than 400 years, journalistic activity has only taken place throughout the last 170 years. This offers an excellent opportunity to observe an emerging genre in an emerging language within an environment that is much closer to the present than medieval stages of language.8 In a brief study, we have analysed three small partial corpora of approximately 10,000 tokens each, focusing on clause linkage structure. They are organised in chronological order:

- From January 1839 we have used the first two issues of the *Grischun Romontsch* (*GrR*), the valley’s first newspaper, founded in 1836, and the *Amitg della Religiu n e della Patria* (*ARP*), founded in 1838.

---

8 For a more detailed vision of the historical context and for further illustration of the argument outlined below, please refer to Obrist (in press).
- From January 1906, we have excerpts from the first two issues of both the *Gasetta romontscha (GaR)*, a conservative newspaper issued for the first time in the 1850s, and from the *Grischun (Gr)*, its progressive counterpart.

- Since there has been a press monopoly in the upper Rhine Valley since World War I, the last diachronic partial corpus we offer is only taken from a single newspaper: *La Quotidiana (LQ)*.

The results, displayed in a similar fashion as those for Old Spanish, are displayed in Table 2.

Table 2: Junction analyses for Rheto-Romance newspapers (syntax)

<table>
<thead>
<tr>
<th></th>
<th>1839</th>
<th>1906</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GrR</td>
<td>ARP</td>
<td>Total</td>
</tr>
<tr>
<td>Coordination</td>
<td>27.3</td>
<td>18.2</td>
<td>22.9</td>
</tr>
<tr>
<td>II-III</td>
<td>57.0%</td>
<td>61.7%</td>
<td>58.6%</td>
</tr>
<tr>
<td>Subordination</td>
<td>10.0</td>
<td>4.5</td>
<td>7.4</td>
</tr>
<tr>
<td>IV</td>
<td>20.9%</td>
<td>15.3%</td>
<td>18.9%</td>
</tr>
<tr>
<td>Gerunds</td>
<td>0.6</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>V</td>
<td>1.3%</td>
<td>3.4%</td>
<td>2.0%</td>
</tr>
<tr>
<td>PrepP</td>
<td>10.0</td>
<td>5.8</td>
<td>8.0</td>
</tr>
<tr>
<td>VI-VII</td>
<td>20.9%</td>
<td>19.7%</td>
<td>20.5%</td>
</tr>
<tr>
<td>Total</td>
<td>47.9</td>
<td>29.5</td>
<td>39.1</td>
</tr>
</tbody>
</table>

The most general observations concerning this corpus are:

- In contrast to the Old Spanish corpus, the overall score of junctors shows a decrease rather than an increase through time.

- The frequency of overt linkage of finite clauses, both coordinated and subordinated, is clearly diminishing, while the prepositional phrases have remained stable or are even
increasing slightly. Comparing the relative weight of junction techniques, we can see that prepositions are becoming more important. This happens at the expense of coordination, but not only that; unlike in Old Spanish, subordination also diminishes with coordination.

- Variation within chronological pairs is, again, considerable. It is interesting to observe that the newer of two newspapers usually anticipates the evolution to come: ARP was not founded until 1839, and Gr in 1905, whereas their counterparts GrR and GaR were already established; however, we cannot claim statistical significance in this respect.

We will seek insight into the statistical evolution using illustrative passages of the underlying texts. The opening article from the international news section of the ARP, for example, begins as follows:

(1) Frontscha. A Paris fuva la generala creta che cun la redunonza dils deputai vignessan las gassas bubrontadas cun saung, ni silmains rumpessi ora ina ravolta. **Denton** ein las combras aviartas senza disturbar strusch ina miur. **Bein zvar che** sez il retg haveva tema de malemperneivlas novadats. **Pertgei** tutt Paris fuva sin quella chischun emplenius cun postas militaras et il retg circumdaus cun extraordinarias excortas. Finalmegn **arrivaus** cun il tgau saun el saal della redunonza, ha el teniu tier ils deputai in plaid de quest contegn: [...] **Aber per** miserabla memoria ha Philipp gnianc tratg a strada las strepitusas demonstratiuns e tiarra-triembels sin ils confins della Svizzera, **sonder lura sin** in solemna vivat dils deputai turnaus cun piall sauna en siu casti.

“In Paris, the general opinion was that the meeting of the Parliament would cause the streets to flood with blood, or at least, that a revolt would break out. **However,** the houses opened without even startling a mouse. **It is true that** even the king was prepared for bad news. **For this reason,** all of Paris was full of military checkpoints on this occasion, and the king [himself] surrounded by an unusual amount of guards. Finally, upon **arriving** in an uninjured state, he addressed the deputies with the following words: [...] **But because** of his bad memory, Philipp did **not even** bring up the loud demonstrations and unrest on the Swiss border; **instead,** he [then] returned safely to his castle with a solemn ‘vivat’ to the deputies.”

The content of the article is, actually, a simple list of facts. The arguments can be summarised as follows: Paris had been fearing a general uprising on the occasion of the parliament meeting. Security measures were taken. The king held a speech without referring to the agitation on the Swiss border; he returned to his palace. However, if we were given only the series of junctors: **denton** (“all the same”)... **bein zvar che** (“it is true that”)... **pertgei** (“for this
reason”)... aber (“but”)... per (“because of”)... gniand/sonder (“not even”/“instead”), our first impression would probably be that this is an editorial article. Indeed, we have what Douglas Biber (1993: 11) calls an overt expression of persuasion: a high frequency of conjunctions expressing causality and similar relations (concessivity is in fact based on causality).

The following article from the first issue of Grischun in 1906 also deals with a situation of crisis. However, in comparison to the former text, its style seems somewhat more modern, if still not quite contemporary to us:

(2) Russia. La situazion ci cheu disparada; eun pauc ploids: Revoluziun e confusiun sin tut la lingia. Ad Odessa han ils consuls dellas differentas naziuns perfin decidiu de tener prompts ils bastiments, per transportar a casa lur compatriots pil cass de basegns. Il consul svizzer ha domondau il cusseg federal, sch’el vegli surprender las spesas de transport de quels Svizzers, che seigin senza miezs. Il cusseg federal ha respondiu en semn affermativ, resalvont il regress vid ils cantuns u las vischnauncas.

“The situation here has exploded: in a few words, revolution and confusion everywhere. In Odessa, the consuls of several countries have finally decided to prepare ships in order to repatriate their citizens in case of necessity. The Swiss consul has asked the government if it is willing to pay for the transport of these Swiss, [alleging] that they are without means. The government has answered affirmatively, reserving [for itself] [the right of] recourse against the cantons and municipalities.”

The differences between (1) and (2) are obvious: the newer article (2) shows less overt linkage and when it uses junctors, the junctors are in the form of noun and infinitive phrases more often than the junctors are in the older article (1): per transporar a casas “in order to bring home,” pil cass de basegns “in case of necessity,” resalvont il regress “reserving the regress.” Furthermore, the underlying semantic relations are not made explicit: in the latter example, the possibility of holding the repatriates responsible for the expenses they have caused does of course establish a relation of condition or contrast with the assistance they are given by the government, but there is no but or although.

Does overt clause linkage drop between 1839 and 1906 because of a more implicit structure of argumentation? A re-classification of the data seems to confirm this intuition. We have split up the coordinating and subordinating junctors into two semantic groups: those
usually considered as descriptive and narrative (e “and,” lura “then,” etc.), and those that show expository or argumentative functions (sche “if,” perquei che “because,” aber “but,” etc.). The results show a more differentiated image: the high overall junctor score for 1839 is indeed based on argumentative functions, including condition, cause, and concession.

Table 3: Junction analyses for Rheto-Romance newspapers (semantics)

<table>
<thead>
<tr>
<th></th>
<th>1839</th>
<th>1906</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GrR</td>
<td>ARP</td>
<td>Total</td>
</tr>
<tr>
<td>Addition &amp; temporal</td>
<td>18.4</td>
<td>10.3</td>
<td>14.5</td>
</tr>
<tr>
<td>of which et</td>
<td>38.4%</td>
<td>34.9%</td>
<td>37.1%</td>
</tr>
<tr>
<td>Argumentative functions</td>
<td>10.0</td>
<td>6.5</td>
<td>8.4</td>
</tr>
<tr>
<td>Total</td>
<td>29.5</td>
<td>19.2</td>
<td>24.6</td>
</tr>
<tr>
<td></td>
<td>61.6%</td>
<td>65.1%</td>
<td>62.9%</td>
</tr>
</tbody>
</table>

A major deviance from the general evolution is the value of additional and temporal junctors in the Gasetta Romontscha. This is mainly due to the junctor e (“and”), the frequency of which is twice the value as of the coetaneous Grischun paper. It is of course legitimate to inquire about the reasons for this important variation. We must insist, however, that this is not a mere issue of corpus design: e-junctors are well distributed throughout the text, and it is impossible to find a passage with an excessive presence of enumerations. If the reason for the variation is idiosyncrasy, the idiosyncrasy must pertain to the style of the newspaper as a whole, and not to that of a particular article. Consider the following example, taken from the local news section:

(3) Als vischins dallas tracziuns de Puschlav vuleva la malvegnida prescripziun de stuer pinar la lena sort da cumminonza buc el igau e han els perquei concludiu de reparter dad els anora la lena marcada ed ein era, encuter el scammond della suprastonza comunala, sepatroni de quella.
“To the residents of the Puschlav villages, the unwelcome prescription of dividing timber among the community did not make any sense, and they therefore decided to distribute the timber of their area by themselves, and they also, against the prohibition of the council, took possession of it.”

The information here is given as a sequence: malcontentedness with the rules set by the authorities, disobedient autonomous repartition, and finally usurpation. There is topic continuity in the sense that the thematic position of all three sentences is occupied by the Puschlav residents. In three of the two sentences, they also make up the grammatical subject. There is a clear structure in this passage; however, the way the information is presented in a style that does not seem to be as close to an objective, journalistic reporting of the facts it could be. Its incremental character (conjunctions followed by adverbs) makes it somehow rhetorically marked.

Consider the local news from 2006, here, in form of an obituary:

(4) Gest il mument ch’il temps da pensiun era arrivaus, ha sia sanadat entschiet a far quitaus. Ina malsogna maligna s’annunzia e s’avonza cun ried. Liungs tractaments medicinals, diversas dimoras en spitals e la buna tgira a casa han bein purtau levgera, mo buc il migliurament ton desiderau.

“At the very moment of reaching the age of retirement, his health began to decline. A malign illness announced itself and proceeded quickly. Long medical treatments, long stays at hospitals, and good care at home brought relief indeed, but not the recovery so hoped for.”

This passage is similar to the precedent in its length (three sentences) as well as in the organisation of its content: someone has reached retirement age, he is diagnosed with cancer, he is given every possible treatment, and he still dies. The way this information is presented is different, however: the uniting semantic feature of the three sentences (and the whole article), namely the patient, is neither featured in thematic function nor is it the grammatical subject. Instead, each sentence begins with a different agent: first the general health condition, then the illness, and finally the treatment. Coherence between the sentences is mainly lexical; there are hardly any overt signals of cohesion, except perhaps for the verb
tense. It is important to observe that (4) can easily be paraphrased in such a way that topic continuity be respected: *Just in the moment of reaching the age of retirement, his health started to decline* (and) *he was diagnosed with a fast growing cancer*. As observed by Kreyer (this volume), a topic shift as in (4) appears to be an indicator in and of itself. It would indeed be interesting to correlate the presence of junctors – especially coordinators – with phenomena such as inversion and dislocation.

For the time being, the findings of this chapter may be summarised as follows. The main goal of newspaper texts is transmitting news, and this has been so ever since newspapers have existed. Still, clause linkage shows considerable diachronic and synchronic variation. First, overt clause linkage tends to decline over the years. Second, this decline affects conditionals and causals more strongly than it affects other semantic areas of clause linkage. Third, there is a diachronic shift from more aggregate to more integrate types of linkage.

Many of the phenomena described in this section have been said to be functional necessities of journalistic writings (see Lüger 1983; Kirstein 1997; Carreira 1999). For example, Eggers (1973: 46) replaces a noun cluster taken from a newspaper article by a series of verb phrases that reach twice the length of the initial structure. He cites this as evidence that noun phrases fit more easily into journalistic language because they allow for more concise formulations (see also Puschmann, this volume). However, it is important to distinguish between textual function and traditionality of discourse. The argument of conciseness does not necessarily apply to examples like (4), where *pil cass de basegns* (“in case of necessity”) could easily be rendered alternatively with *sche quei fuss necessari* (“if this is necessary”). As well, apart from a very modest contraction of the text, there is no need to avoid causal markers and series of *e*-junctors.
From a synchronic point of view, it can be argued that newspaper language, as we describe it, obeys conventionalised patterns or discourse traditions. From a diachronic point of view, we can retrace these patterns to their origins and describe their evolution. An important terminus in this respect is genre differentiation, as described by Raymund Wilhelm (2001: 472), a fundamental mechanism in the creation of new discourse traditions. Many of the nineteenth-century newspaper articles we analysed resemble essays or sermons. Wilhelm’s research on the origins of journalism in Italy shows that these are precisely the genres which served as models for emerging journalism. If it is true that today's commentaries and lead articles still show some affinity to older journalistic language, as we have stated with respect to example (1), we might even postulate that commentaries follow a more archaic discourse tradition than ordinary news texts.9

7. Conclusions

The question we have discussed in this paper is to what extent textual traditions, in the sense of discourse tradition theory, have to do with patterns of clause-linking elements. The initial hypothesis, that clause linkage might be a strong indicator of different textual traditions, was supported by the traditional view on textual elaboration and by our previous studies where a strong correlation between the two was found in some Old Spanish texts. We therefore argued that not only different clause-linking techniques, but also their relative frequencies, are symptomatic of different textual traditions, and this is so not only in the context of genres but also in the context of sub-tradition(s) within a single genre.

In the initial design of our study, we aimed to create global descriptions of the clause-linkage patterns found in each text and we applied this process to three corpora: a diachronic corpus of Old Spanish texts including different genres, an Old Spanish parallel corpus of

9 An interesting formal mark in this respect is that the renowned Frankfurter Allgemeine Zeitung still prints the titles of its commentaries in Gothic letters, while normal serifs are used for all other headings.
Bible texts, and a Rheto-Romance diachronic newspaper corpus from the nineteenth and twentieth centuries.

It is obvious that texts are complex entities and that any reductive projection of text properties, even a bi-dimensional and elaborated one such as the one that was proposed in our study, necessarily requires that strong limitations be imposed upon it. But instead of giving up on looking for criteria that can be formalised and limiting ourselves to purely philological descriptions of texts as incomparable individuals, we believe that formalisations are necessary, and this is perhaps even more the case for purposes other than textual typology alone. Textual typology, in the sense of a self-sufficient discipline, can work with complex criteria to define the individual properties of a text, but if we think, for example, that corpus linguistics is interested in genre differentiation in order to adequately describe grammatical variation in a language or the evolution of grammar in diachrony, we need reductive criteria that also allow for a comparison of larger amounts of texts. Nevertheless, and even if clause-linkage patterns are strong indicators of certain textual traditions, other features that might be even more indicative of the differentiation of certain text traditions should also be taken into account. On the other hand, in some cases it seems clearly that textual traditions can be identified via clause-linkage analysis. In the case of the Bible corpus, it was even possible by objective measure of clause-linking elements to confirm the philological reconstruction of similarity and difference between different versions, and the junctogram analysis allowed, in some of the Bible examples, for the distinguishing of factors of textual progression from individual deviances between the different versions. Obviously, clause-linkage analysis also allows for the evolution of clause-linkage techniques themselves when applied to a diachronic corpus that is limited to one genre. In this way, the evolution of the Rheto-Romance newspapers showed the Ausbau of this genre from a more aggregative to a more integrative style in the last two centuries.
What our results perhaps more clearly show are the possible achievements and limitations of the study of junction patterns for historical linguistics. For more general observations on the significance of clause-linkage patterns for genre characterisation (or, in a more general way, for the identification of discourse traditions), we will need to study a larger number of texts from different languages, not only from Romance languages and not only in a diachronic perspective, and we will have to compare the clause-linkage-patterns with other linguistic features.

It seems that the tendency in linguistics to simplify or even to avoid the question of genre differentiation has to do with the enormous complexity of textual traditions. Corpus linguistics deals with an ever-increasing mass of data and sometimes this simple mass allows for important observations. However, the analysis of a large data mass also creates new questions, and since the mass only allows for a first look at the complex and differentiated situation or evolution behind the data, we absolutely need criteria for the objective analysis of the discourse traditions of which our corpora are composed. This differentiation will not only lead to an atomistic view that does not allow for more general states or evolutions to be seen, but it will contribute to an adequate conception of what a language really is and how some phenomena are text specific or discourse tradition specific, while others affect the language as a whole. This is surely the most important contribution the notion of discourse traditions can make to linguistic theory.

8. References

Text corpora

Bibles:

Aj = Biblioteca de Ajuda 52-xii-1 [manuscript]

Alb = Biblia de Alba, Madrid, Palacio de Liria [manuscript]
E3 = Escorial I.i.3 [manuscript]
E4 = Escorial I.i.4 [manuscript]
E7 = Escorial I.i.7 [manuscript]
E8 = Escorial I.i.8 [manuscript]
E19 = Escorial I.ii.19 [manuscript]
Faz = Fazienda de Ultramar, Salamanca, Universidad 1997 [manuscript]
GE 1= Madrid, Biblioteca Nacional, 816 [manuscript]

Rheto-Romance newspapers:


ARP = Amitg della Religiuon e della Patria. Surrhein/ Somvitg, 1838-1839 [excerpt: 7-21 Jan 1839]


Gr = Il Grischun. Cuera, 1905-1916 [excerpt: 5-12 Jan 1906]


Other references:


Enrique-Arias, Andrés 2006 Sobre el parentesco entre la *Biblia de Alba* y la *Biblia* de la Real Academia de la Historia Ms. 87. *Romance Philology* 59: 241–263.


Marta Fernández Alcaide and Araceli López Serena (eds.) 2007 *Cuatrocientos años de la lengua del Quijote.* Sevilla: USE.

Frank, Barbara, Thomas Haye, and Doris Tophinke (eds.) 1997 *Gattungen mittelalterlicher Schriftlichkeit.* Tübingen: Narr.


Kreyer, Rolf 2010 Syntactic constructions as a means of spatial representation in fictional prose. In: Heidrun Dorgeloh and Anja Wanner (eds.): Syntactic variation and genre, xxx-xxx. (Topics in English Linguistics xx.) Berlin/New York: Mouton de Gruyter.


Puschmann, Cornelius 2010 “Thank you for thinking we could”, Use and function of interpersonal pronouns in corporate web logs. In: Heidrun Dorgeloh and Anja Wanner (eds.): Syntactic variation and genre, xxx-xxx. (Topics in English Linguistics xx) Berlin/New York: Mouton de Gruyter.


