

Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features

Marina Santini

ITRI

University of Brighton

Lewes Rd, Brighton, UK

Marina.Santini@itri.brighton.ac.uk

Abstract

In this report we propose a new set of features for automatic genre and text type identification: linguistic facets. The label linguistic facet has been created to stress the fact that each of the features in this new set highlights a facet, i.e. an aspect in the communicative context that is reflected in the use of language. Linguistic facets subsume two set of features: functional cues and syntactic patterns. Functional cues are not completely new. Syntactic patterns, instead, have never been tried before in any automatic approach. Both functional cues and syntactic patterns are linguistically-motivated features that can be interpreted functionally. The effort is to derive from text-internal linguistic cues the communicative context in which the text has been produced/consumed. With linguistically-motivated and functionally-interpretable features it is possible to combine qualitative analysis and quantitative findings, and get a more accurate text analysis together with genre and text type identification. Preliminary results show that linguistic facets have a robust discriminating power for web genre classification.

1	INTRODUCTION	3
2	FEATURES USED SO FAR IN AUTOMATIC GENRE AND TEXT TYPE IDENTIFICATION	4
3	RESOURCES: EARLIER LINGUISTIC WORK AND PARSER	5
3.1	LINGUISTIC FEATURES AND THEIR INTERPRETATION.....	5
3.2	THE PARSER.....	7
4	LINGUISTIC FACETS	8
4.1	LINGUISTIC FACETS (I): FUNCTIONAL CUES.....	8
4.2	LINGUISTIC FACETS (II): SYNTACTIC PATTERNS.....	18
4.2.1	<i>Methodology: Creation or Detection of Syntactic Patterns</i>	19
4.3	ADVERBIAL CLAUSES	24
4.4	COMPLEMENT/NOMINAL CLAUSES.....	33
4.5	SIMPLE SENTENCES	37
5	PRELIMINARY RESULTS: ACCURACY OF LINGUISTIC FACETS FOR WEB GENRE CLASSIFICATION	38
6	REFERENCES	40

1 Introduction

In this report we propose a new set of features for automatic genre and text type identification: linguistic facets.

Several definitions of the notions of genre and text type have been proposed so far for an automatic approach (see Santini 2004a for a review). Here by *genre* we refer to the socio-cultural, textual and linguistic conventions used in a document. When conforming to a set of well-recognised conventions, a document becomes a member of recognized family of similar documents, i.e a genre. By *text type*, instead, we refer to the discoursal/rhetorical strategies enacted in the document. According to this distinction, an *academic paper* is a genre with its own textual organization and linguistic conventions, accepted and recognized by academic people, who represent the socio-cultural environment in which this genre is used and useful. Academic papers usually have an *argumentative* text type, which means that the predominant rhetorical strategy enacted in this genre is to “make an argument”, i.e. to demonstrate that the claim put forward by the writer(s) is convincing.

An element that the notions of genre and text type share is that, broadly speaking, they are both orthogonal to the notion of *topic* (Meyer zu Eissend and Stein 2004, Lee and Myaeng 2004, Lim et al. 2005, etc.) In fact, the same topic can be expressed in different genres, using different text types. While the automatic classification of topical categories (text categorization) is based on content words (the so-called bag-of-words approach), the automatic identification of genre and text type is based on grammatical features, ranging from part-of-speech (POS) tags, to functional words, some ratios, syntactic cues, and so on (see next section). The broad term *identification* is used here as an umbrella term subsuming analysis, detection and classification. While the automatic approach to text type has focussed on textual analysis, language variation and detection (Biber 1988, 1995, Nakamura 1993, Wickberg 1993, etc.), the automatic approach to genre has centered upon classification tasks (Cutting and Karlgren 1994, Kessler 1994, etc.). Both automatic approaches tend to share the same family of features.

The label *linguistic facet* has been created to stress the fact that each of the features in this new set highlights a facet, i.e. an aspect in the communicative context that is reflected in the use of language. For example, the use of activity verbs (see Subsection 4.1), which denote actions related to choices, is very common in instructions and conversation. Therefore the facet “activity verbs” links the frequencies of occurrence of a specific group of verbs in a text to the purpose of the text (to give instructions) or to the live situation (a conversation). The frequencies of occurrence, or sometimes the mere presence of a facet, help us understand the use and the variation of the language across different kinds of texts.

Linguistic facets subsume two set of features: *functional cues* and *syntactic patterns*. Functional cues are not completely new. Syntactic patterns, instead, have never been tried before in any automatic approach. Both functional cues and syntactic patterns are linguistically-motivated features that can be interpreted functionally. The aim of a functional interpretation is to link linguistic features to the communicative situation which represents the external context. The effort is to derive from text-internal linguistic cues the communicative context in which the text has been produced/consumed. The part of the communicative context we are interested in is the one related to genre and text type. With linguistically-motivated and functionally-interpretable features it is possible to combine qualitative analysis and quantitative findings, and get a more accurate text analysis together with genre and text type identification.

Another advantage of linguistic facets is their corpus-independence. A drawback of most of the automatic feature selection methods, -- when they work, cf. for example Copeck et al. (2000) -- is that they return subsets of features which are tailored to the corpus from where the features were extracted. Automatically selected features often "overfit", which means that they are well-suited for the source corpus, but they cannot be easily exported to a corpus of different genres or text types, because their representativeness is lost (for an experience, see Santini 2005). Linguistic facets overcome this problem. As they come from grammar books and represent common grammatical features, it is the variation in their distribution, or their mere presence, that gives hints about their use across different kinds of texts.

It is important to remind that the set of linguistic facets suggested here is only an initial set that can be easily enlarged and enhanced in future.

The report is organized as follows: Section 2 gives account of features used so far in genre and text type identification; Section 3 describes the resources used for the creation of linguistic facets; Section 4 provides a detailed breakdown of linguistic facets; Section 5 reports preliminary results of the performance of linguistic facets in a classification task; Section 6 suggests the direction for future work.

2 Features Used So Far in Automatic Genre and Text Type Identification

Only very few studies specifically comment on features which are useful for automatic genre and text type identification.

In the mid-80s Biber made a huge effort to collect 67 linguistically-motivated features, mainly from socio-linguistic studies and from Quirk et al. (1985). A thick Appendix in his *Variation across speech and writing* (Biber 1988) is dedicated to the description of these features, which were used with the multi-

dimensional approach, i.e. a quantitative, corpus-based approach based on factor analysis. Biber's methodology is geared towards text type identification

Wolters and Kirsten (1999) investigated to what extent POS frequencies and content words are useful for both domain and genre classification in a German corpus. But their results were inconclusive. They suggested that the use of these two sets of features should be investigated separately.

Dewdney et al. (2001) compared the performance of presentation features (linguistic features and layout features) and word features (bag-of-words approach) using three different classifiers (Naïve Bayes, C4.5 and SVM) over the Carnegie Mellon University (CMU) genre corpus. According to their results, the use of presentational features aided by some selected words is the most promising for genre classification.

Following Argamon (1998), Santini (2004b) used POS trigrams to discriminate across ten genres in the BNC. A Naïve Bayes classifier reached 87% accuracy across ten genres -- *conversation, interview, public debate, planned speech, academic prose, advert, biography, instructional, popular lore, and reportage*.

Other features were used for automatic genre and text type identification, namely:

- specific verbs (Nakamura 1993, Wickberg 1993); relativisers (Sigley 1997) for text type identification;
- POSs (Karlgrén and Cutting 1994); lexical cues, character level cues, ratios (Kessler et al. 1997); most frequent words in English and punctuation (Stamatatos et al. 2000) for automatic genre classification.

It is almost impossible to compare the results of these studies and draw any general conclusions about the best set of features, because different corpora, different statistical approaches, different sets of genres and text types were used all the time.

3 Resources: Earlier Linguistic Work and Parser

Linguistic facets are based on two main resources: earlier linguistic work and a NLP tool. The theoretical side of linguistic facets is based on four grammars (Quirk et al. 1972; Werlich 1976, Quirk et al. 1985, Biber et al. 1988), and a corpus-based study on language variation (Biber 1988). The empirical side is based on the use of the parser Connexor (Tapanainen and Järvinen 1997). The motivation of the selection of these resources is given in the following subsections.

3.1 Linguistic Features and their Interpretation

The set of linguistic facets suggested here includes only features that can be automatically extracted from a text. The design of linguistic facets relies on earlier linguistic works, namely: Werlich (1976), heavily based on Quirk et al.

(1972); Quirk et al. (1985); Biber (1988), which draws many features from Quirk et al. (1985); and Biber et al. (1999). These works have something in common: they provide an interpretation of the features according to different kinds of texts. Werlich (1976), Biber (1988) and Biber (1999) make this intent explicit. In Quirk et al (1972) and Quirk et al. (1985), instead, reference to the preferred use of some linguistic phenomena in different kind of texts is more occasional.

Werlich belongs to the German tradition of *textlinguistik*. His text grammar is a descriptive grammar, where claims are supported by examples from real texts, but it is not corpus-based (the corpus-based approach was not widespread in the mid-70s, especially outside the UK). As mentioned earlier, the linguistic and textual features described in Werlich (1976) are mainly taken from Quirk et al. (1972). The added value by Werlich is the description of the use of these features across five text types: *description*, *narration*, *exposition*, *argumentation* and *instruction*. He uses a somewhat peculiar terminology, not linked to the English-speaking linguistic tradition, and hypothesizes "five dominant contextual foci that can be observed in all texts" (Werlich 1976: 19): a spatial focus or context for descriptive texts; a temporal focus or context for narrative texts; a focus on concepts for expository texts; a focus on relations between concepts for argumentative texts; a focus on future behaviour for instructional texts. His text grammar is based on the hypothesis that texts are "conceived of as assignable to text types [and] primarily derive their structural distinctions from innate cognitive properties" (Werlich 1976: 21). In his view, the five basic text types reflect the basic cognitive processes of contextual organization: description relates to *perception in space*; narration is linked to *perception in time*; exposition depends on the *comprehension of general concepts*; argumentation relies on *judging*; instruction is based on *planning*.

The contribution of Werlich's text grammar is that it provides breakdowns of linguistic and textual features for each text type. These features have the potential to be useful for any automatic genre and text type identification. Only a restricted set of features have been drawn from Werlich's grammar, namely the phenomenon sentences (see Subsection 4.5), i.e. simple sentence patterns specific to the five text types.

While Werlich's text types derive from a cognitive interpretation of texts in relation to text producer, text receiver(s), external context and so on, Biber's text types are corpus-based, statistically extracted with factor analysis and interpreted in terms of "textual dimensions" (Biber 1988: 104 ff.). From a corpus of 23 genres, he computes the frequencies of 67 linguistically-motivated features, derived from sociolinguistic studies and from Quirk et al. (1985), and extracts 7 factors. He suggests interpretive labels for each factor (except Factor 7, which seems to be not sufficiently representative) and ends up with 6 dimensions: *informational vs. involved production*; *narrative vs. non-narrative concerns*; *explicit vs. situation-dependent reference*; *overt expression of persuasion*; *abstract vs. non-abstract information*; *online informational elaboration*. His text types cut across genres (Biber 1988: 129 ff.). They are statistically validated by a

confirmatory cluster analysis and claimed to be representative for the English language (Biber 1989). Nowadays they look too corpus-dependent and too subjective to be representative for English. However, the linguistic analysis brought about by Biber's work remains a remarkable contribution to linguistics.

Biber (1988) dedicates a whole chapter (Chapter 2) to the communicative functions served by linguistic features. In a useful table, it outlines all the functions related to linguistic features (Biber 1988: 35). This functional interpretation of linguistic elements is entirely incorporated into Biber et al. (1999: 41-44). Biber et al. (1999) (with its useful student book) is entirely corpus-based. The LSWE Corpus (the Longman Spoken and Written English Corpus) contains over 40 million words of text (Biber et al. 1999: 24). It has been constructed to provide a representation of different registers, namely: *conversation, fiction, news, and academic prose*. Each grammatical feature is described across the four registers following a regular structure: first a description of the grammatical feature in question; then a section called 'Corpus Findings' where the distributional patterns of the grammatical features are shown; finally a section called 'Discussion of the Findings' where the discourse patterns described quantitatively are illustrated and interpreted in functional terms (Biber et al. 1999: 41). This grammar represents a linguistic resource to be mined for decades.

Quirk et al. (1972) is used here mostly in an indirect way, but it is important because it is the foundation of Werlich (1976).

Quirk et al. (1985) is an invaluable descriptive grammar for English, which is only partially corpus-based (Quirk et al. 1985: 33). It is interspersed by remarks about formality and other situational dimensions reflected into the language. For example, for a construction such as the "verbless temporal clause with *until* in final position", as in *Beat the mixture until fluffy*, is "common in instructional texts" (Quirk et al. 1985: 1079).

3.2 The parser

In proposing linguistic facets for genre and text type identification, we wish to take full advantage of the linguistic knowledge returned by a parser. Ideally, with our approach many different parsers outputting complementary information could be used for analysing the same text. In this first version of linguistic facets, we only use one parser, i.e. Connexor by Tapanainen and Järvinen (1997).

Connexor is based on a functional grammar, and outputs the baseform, or lemma, and four kinds of annotation: syntactic relations, functional tags (starting with @), syntactic tags (starting with %), and morphological tags, which do not show any special prefix.

This is an example of the output of the sentence *The cat is on the mat*:

#	Text	Baseform	Syntactic relation	Syntax and morphology
1	the	the	det:>2	@DN> %>N DET
2	cat	cat	subj:>3	@SUBJ %NH N NOM SG
3	is	be	main:>0	@+FMAINV %VA V PRES SG3
4	on	on	loc:>3	@ADVL %EH PREP
5	the	the	det:>6	@DN> %>N DET
6	mat	mat	pcomp:>4	@<P %NH N NOM SG
7	.	.		
8	<s>	<s>		

Figure 1. A parsed sentence from Connexor

4 Linguistic Facets

The following subsections describe the initial set of linguistic facets for genre and text type identification: 41 functional cues and 25 syntactic patterns. Both subsets are extracted from the output of Connexor, except most of the connectives, which are drawn from unparsed text.

4.1 Linguistic Facets (I): Functional Cues

Functional cues are mainly based on functional tags (i.e. tags starting with an @ sign according to Connexor annotation scheme, for example @SUBJ), to a minor extent on syntactic relations (for example, loc:> or tmp:>), and on a number of base forms or lemmas (for example, the base forms of semantic classes of verbs). Connectives are extracted from unparsed text and ambiguities are solved with the help of punctuation marks.

The list of functional cues included in this first set of linguistic facets is shown in Figure 1. For each of them, a short description is provided together with the annotation tags or base forms used in the extraction process:

1. *predicators*
2. *complex NPs*
3. *nominals*
4. *first person*
5. *second person*
6. *third person*
7. *third person singular inanimate*
8. *present tense group*
9. *past tense group*
10. *imperatives*
11. *active*
12. *passive*
13. *time markers*
14. *location markers*

15. *instrument markers*
16. *manner markers*
17. *negative particles*
18. *probability markers*
19. *necessity markers*
20. *existential there*
21. *expressiveness*
22. *activity verbs*
23. *communication verbs*
24. *mental verbs*
25. *causative verbs*
26. *occurrence verbs*
27. *existence verbs*
28. *aspectual verbs*

29. enumerative connectives
 30. equative connectives
 31. reinforcing connectives
 32. summative connectives
 33. appositive connectives
 34. resultative connectives
 35. inferential connectives

36. reformulatory connectives
 37. replacive connectives
 38. antithetic connectives
 39. concessive connectives
 40. discursal connectives
 41. temporal connectives

Figure 2. List of functional cues.

Predicators. This facet includes verbs in any roles, from auxiliary to finite or non-finite forms, to -ing forms. An extensive use of verbs is usually connected to an interactive and affective purpose (Biber 1988: 105). In particular, verbs are extremely common in conversational genres (such as interviews or debates), because spoken genres often have a high number of short sentences, to keep the listeners' attention alert. Usually the level of syntactic complexity is very low since each sentence expresses a single main idea. Verbs are also used to show the speaker's attitudes, with elements such as: *I think*, *I know*, etc. (Biber et al. 1999: 359-360).

@+FAUXV # finite auxiliary predicator
 # *Ex: This will cause a soft boot in the emulator.*

@-FAUXV # nonfinite auxiliary predicator
 # *Ex: Software can be split roughly into two main types.*

@+FMAINV # finite main predicator
 # *Ex: Sockets form the interface between these two components.*

@-FMAINV # nonfinite main predicator
 # *Ex: If you are running the DOS emulator, go to the main window.*

@<P-FMAINV" # nonfinite clause as preposition complement
 # *Ex: This mechanism is used for creating a virtual connection.*

Complex NPs. The basic structure of the noun-headed phrase includes four major components, two of which are optional: determiner + (pre-modification) + head noun + (post-modification and complementation). Overall, noun phrases with a modifier are relatively rare in conversational genres, while they are common in press genres and academic genres (Biber et al. 1999: 574 ff.).

@SUBJ who@SUBJ
Ex: anyone who is willing to listen.

@NH _that@SUBJ
Ex: a market system that has no imperfections.

@NH _that@OBJ @SUBJ
Ex: a small wooden box that he owned.

@NH _that@CS @SUBJ
Ex: the fact that I haven't succeeded.

@NH for@<NOM
Ex: the new training college for teachers.

@NH \$s_v_main
Ex: that job I was doing last night.

@NH.*?@INFMARK>

Ex: enough money to buy proper food.

@NH FMAINV_ING

Ex: the imperious man standing under the lamppost.

@NH @ADVL

Ex: a block behind.

@NH.*?@<NOM

Ex: doctors at the John Hopkins Medical School.

FMAINV_ING @OBJ

Ex: detecting devices.

@DN> _that@DN> @NH

Ex: both those copies.

@A> @SUBJ

Ex: a stationary element held in position by the outer casing.

@A> @APP

Ex: the big one in town.

@A>.*?@NH)

Ex: the industrially advanced countries.

Nominals. This facet includes nouns, prepositional phrases, adjectives, appositions, and other elements that expand or enrich the meaning expressed by the noun. High frequency of nouns indicates density of information (Biber 1988: 105). Prepositional phrases integrate other pieces of information into a text, while adjectives and other elements further elaborate nominal information. The nominal component is very high in academic genres (Biber et al. 1999: 235) and informational texts (Biber 1988: 104).

@APP

Apposition

Ex: John, your brother, came yesterday.

@NH

Stray noun phrase

Ex: The house beside the church.

@A>

Premodifier of a nominal

Ex: These sophisticated algorithms are stored safely.

@DN

Determiner

Ex: The optical laser is a useful tool for eye treatment.

@<NOM-OF

Postmodifying prepositional phrase beginning with "of"

Ex: This will cause a soft boot of the emulator.

@<NOM

Postmodifier of a nominal

Ex: This mechanism is used for creating a connection between the two components.

@<P

Other preposition complement

Ex: This will leave the rest of the system running.

@OBJ

Object

Ex: This will leave the rest of the system running.

@SUBJ # Subject
Ex: Software can be split into several categories.

@PCOMPL-S # Subject complement
Ex: Biscuits which are sugar-free, or nearly so.

@PCOMPL-O # Object complement
Ex: This procedure makes the flow transparent to the user.

@QN> # Premodifying quantifier
Ex: This process takes less than 1000 milliseconds.

@VOC # Vocative
Ex: Just do it, John!

First person. This facet comprises first person singular and plural pronouns, including possessives and reflexives. When there is a high frequency of first person pronouns in a text, the communication context appear to be related to the text producer(s) (Werlich 1976: 135, 137). First person pronouns are used in all subjective genres such as comments and opinions.

PRON PERS SG1 # Personal pronouns, singular
I, me, myself, my, mine.

PRON PERS PL1 # Personal pronouns, plural
we, us, ourselves, our, ours.

Second person. This facet comprises second person pronouns, including possessives and reflexives. When there is a high frequency of second person pronouns in a text, the communication context appear to be related to the text receiver(s) (Werlich 1976: 136). Second person pronouns are used in instructional genres and in conversational genres.

<i>you</i>	<i>your</i>
<i>yourself</i>	<i>yours</i>
<i>yourselves</i>	

Third person. This facet comprises third person singular and plural pronouns, including possessives and reflexives. When there is a high frequency of third person pronouns in a text, the communication context appear to be related to persons in the spatio-temporal context outside the producer-receiver communication process (Werlich 1976: 136). Third person pronouns are widespread in all genres (Connexor does not have a tag for third person pronouns).

<i>he</i>	<i>his</i>
<i>she</i>	<i>hers</i>
<i>him</i>	<i>they</i>
<i>her</i>	<i>them</i>
<i>himself</i>	<i>themselves</i>
<i>herself</i>	<i>theirs</i>

Third person singular inanimate. This facet comprises the pronoun *it*, including possessives and reflexives (Werlich 1976: 137). Third person pronouns are used in all genres and often indicate an objective or impersonal point of view.

```

it subj:> # it in the subject position
          # Ex: It strikes me.
          # but not Make it softer.

its
itself

```

Present tense group. This facet focuses on present tense, but comprises also present perfect and future. The present group is usually used in description, exposition, argumentation and instruction (Werlich 1976:144). It is less common in narration.

```

@+FMAINV %VA V PRES # Simple present
                    # Ex: It rains; I usually eat salad; etc.

have main:>0 @+FMAINV %VA V PRES # have as main verb
                    # Ex: I have lunch with him.

had.*?@+FAUXV %AUX V # Present perfect
                    # Ex: I have developed a system;

be main:>0 @+FMAINV %VA V PRES # be as copula
                    # Ex: She is nice; He is a sailor; etc.

will.*?@+FAUXV %AUX V AUXMOD # Future tense
                    # Ex: You will see her later.

shall.*?@+FAUXV %AUX V AUXMOD # Future tense
                    # Ex: I shall go tomorrow.

```

Past tense group. This facet focuses on past tense, but comprises also present and past perfect (Werlich 1976: 144). Narrative texts show high frequencies of verbs in the past group.

```

@+FMAINV %VA V PAST # Simple past
                    # Ex: I visited him yesterday.

had.*?@+FAUXV %AUX V # Past perfect
                    # Ex: I had developed a system;

```

Imperatives. This facet includes only imperatives. According to Biber et al. (1999), the frequent use of imperatives is more common in conversation than in writing. The frequent use of imperatives in conversation is due to the fact that the situation is interactive, "with participants often involved in some sort of non-linguistic activity at the moment of speaking. In such situations, it is natural to use language for the purposes of monitoring the actions of the addressee" Biber et al. (1999: 221). Imperatives are also widespread in instructional genres (Werlich 1976: 265), and persuasive genres.

```

%VA V IMP # Imperative
          # Ex: Eat it up!

```

Active. The active voice basically presents changes. The subject is usually viewed as an agent, affecting the external context. (Werlich 1976: 147). The active voice is the normal choice in many genres (Biber 1999: 475 ff).

```

%VA # Active voice

```

Ex: We describe the experiment in following section.

Passive. In the passive voice, the subject is usually viewed as a recipient affected by something coming from the external context. It can also be used in objective genres to introduce impersonal third person point of view. In some genres, the passive voice is frequently used for preserving anonymity (for example, in instructions) or to suggest general validity for opinions (for example, in comments) (Werlich 1976: 148).

%VP # *Passive voice*
 # *Ex: The experiment is described in following section.*

agt # *Agent*
 # *Ex: The dog was chased by the boys.*

Time markers are connected to narrative genres (Werlich 1976: 19, 39).

dur:> # *Duration*
 # *Ex: This was done in the last 25 years.*

frq:> # *Frequency*
 # *Ex: It often involves the use of an additional tool.*

tmp:> # *Time*
 # *Ex: When you leave, shut the door.*

Location markers are connected to descriptive genres (Werlich 1976: 19, 39).

loc:> # *Location*
 # *Ex: This was done in the USA.*

sou:> # *Source*
 # *Ex: He will move from home very soon.*

pth:> # *Path*
 # *Ex: He travelled from Stockholm to Rome.*

Instrument marker indicates the items used to undertake a task. It is connected with instrumental genres (Werlich 1976: 268).

ins:> # *Instrument*
 # *Ex: He sliced the bread with an electric knife.*

Manner marker indicates the way in which something is done. It is connected to descriptive and instrumental genres (Werlich 1976: 268).

man:>. *?@ADVL # *Manner*
 # *Ex: He cooks poorly.*

Negative particles. Negative forms are much more common in conversation than in writing for several reasons. For example, verbs are more widespread in conversational genres and negation is often linked to verbs (Biber et al. 1999: 159). Negative particles

lose	produce	send	try
make	provide	shake	turn
meet	pull	share	use
move	put	show	visit
obtain	reach	sit	wait
obtain	receive	smile	walk
open	reduce	smile	watch
pass	repeat	spend	wear
pay	run	stare	win
pick	save	take	work
play	sell	throw	

Communication verbs can be considered a special subcategory of activity verbs that involves communication. Communication verbs are common in all genres including reported speech, dialogue or conversation (Biber et al. 1999: 362). The list of verbs included in this facet comes from Biber et al. (1999: 368).

accuse	declare	persuade	speak
acknowledge	demand	phone	specify
address	deny	pray	state
admit	describe	promise	suggest
advice	discuss	propose	swear
announce	emphasise	publish	talk
answer	encourage	question	teach
appeal	excuse	quote	tell
argue	explain	recommend	thank
ask	express	remark	threaten
assure	inform	reply	urge
call	insist	report	warn
challenge	invite	respond	welcome
claim	mention	say	whisper
complain	note	shout	write
consult	offer	sign	
convince	offer	sing	

Mental verbs denote a wide range of activities and states experienced by humans. They do not involve physical action and do not necessarily entail volition. Their subject often has the semantic role of recipient. They include both cognitive meaning and emotional meaning. They are common in conversational genres (Biber et al. 1999: 362, 366). The list of verbs included in this facet comes from Biber et al. (1999: 368).

accept	consider	find	listen
afford	count	forget	love
agree	dare	forgive	mean
appreciate	decide	guess	mind
approve	deserve	hate	miss
assess	detect	hear	need
assume	determine	hope	notice
bear	discover	identify	observe
believe	dismiss	ignore	perceive
blame	distinguish	imagine	plan
bother	doubt	impress	predict
calculate	enjoy	intend	prefer
care	examine	interpret	pretend
celebrate	expect	judge	prove
choose	experience	justify	read
compare	face	know	realise
conclude	fear	learn	realize
confirm	feel	like	recall

reckon	satisfy	suppose	want
recognise	see	suspect	wish
regard	solve	think	wonder
remember	study	trust,	worry
remind	suffer	understand	

Causative verbs indicate that an entity brings about a new state of affairs. They are quite common in academic genres (Biber et al. 1999: 363, 366). The list of verbs included in this facet comes from Biber et al. (1999: 369).

help	affect	force	influence
let	cause	prevent	permit
allow	enable	assist	
require	ensure	garantee	

Occurrence verbs report events that occur apart from any volitional activity. They are quite common in academic genres (Biber et al. 1999: 364, 366). The list of verbs included in this facet comes from Biber et al. (1999: 369).

become	develop	increase	shine
happen	occur	last	sink
change	arise	rise	slip
die	emerge	disappear	
grow	fall	flow	

Existence verbs report a state that exists between entities. Some of the most common verbs of existence or relationship are copular verbs, such as *seem* and *appear*. They are quite common in academic genres (Biber et al. 1999: 364, 366). The list of verbs included in this facet comes from Biber et al. (1999: 369).

appear	illustrate	matter	reveal
concern	imply	owe	seem
constitute	include	own	sound
contain	indicate	possess	stand
define	involve	reflect	stay
deserve	lack	relate	suit
exist	live	remain	tend
fit	look	represent	vary

Aspectual verbs characterize a stage or progress of some other event or activity (Biber et al. 1999: 364). The list of verbs included in this facet comes from Biber et al. (1999: 369).

start	begin	end
keep	continue	finish
stop	complete	cease

Connectives are conjunctive adverbs not integrated in the clause structure that indicate a linkage between what has just been said and what was said before (Werlich 1976: 201). According to Werlich, different sequences of conjunctive adverbs mark different text types (Werlich 1976: 167, 202). For example, appositive, reformulatory, discoursal and

reinforcing connectives are used in expository and instructional genres. Here the different classes of connectives have been taken from Quirk et al. (1985: 634 ff.). A description of their roles in different kind of text can be found in Werlich (1976: 167-179).

Enumerative connectives:

finally	in the second	on the other	thirdly
first	place	hand	three
first of all	last	one	to begin with
firstly	last of all	second	to conclude
for a start	lastly	second of all	to start with
in the first	next	secondly	two
place	on the one	then	
	hand	third	

Equative connectives:

by the same	correspondingly	in the same	likewise
token	equally	way	similarly

Reinforcing connectives:

above all	furthermore	on top of it	to cap it all
again	in addition	on top of it	to top it
also	in particular	all	to top it all
besides	more	then	what is more
further	moreover	to cap it	

Summative connectives:

altogether	in conclusion	therefore	to summarize
all in all	in sum	thus	
in all	overall	to sum up	

Appositive connectives:

for example	in other words	specifically	that is to say
for instance	namely	that is	

Resultative connectives:

accordingly	consequently	now	somehow
as a consequence	hence	of course	therefore
as a result	in consequence	so	

Inferential connectives:

else	in that case	then
in other words	otherwise	

Contrastive (reformulatory) connectives:

alias	better	more	more precisely
alternatively	in other words	accurately	rather

Contrastive (replacive) connectives:

again	on the other	worse
better	hand	

Contrastive (antithetic) connectives:

by comparison	by way of	in comparison	on the
by contrast	contrast	in contrast	contrary
by way of	contrariwise	instead	oppositely
comparison	conversely		

Contrastive (concessive) connectives:

admittedly	at the same	in any event	of course
after all	time	in spite	only
all the same	besides	in spite of it	still
anyhow	besides	all	still and all
anyway	else	in spite of	that said
anyways	for all that	that	though
at all events	however	nevertheless	yet
at any rate	in any case	nonetheless	
		notwithstanding	

Transitional (discoursal) connectives:

by and by	by and bye	by the way	incidentally
-----------	------------	------------	--------------

Transitional (temporal) connectives:

eventually	in the	meanwhile
in the	meanwhile	originally
meantime	meantime	subsequently

4.2 Linguistic Facets (II): Syntactic patterns

The idea that certain genres or writing styles favour certain syntactic constructions is not new (Biber 1988: 229-230; Baayen et al. 1996; Stamatatos et al. 2001, etc.). However, even if syntax is acknowledged to have discriminating power, (though sometimes reluctantly Aaronson 1999), it has often been neglected in genre categorization studies, because the extraction of syntactic features is considered to be computationally expensive and time-consuming (Kessler et al. 1997). The 67 linguistic features selected by Biber more than 15 years ago (Biber 1988: 73-75, 221-245) are based mainly on word identification, even when the features are really syntactic, because NLP tools were quite limited at that time. For example, the identification of adverbial clauses is based on the presence of specific subordinators, such as *although* and *though* for concessive clauses, and *because* for causative clauses. However, the lexically-based approach to syntax is

quite limited, because subordinators can be ambiguous. To overcome the ambiguity issue, Biber used only unambiguous subordinators; for example *because* is the only causative subordinator included in his features, being the only one "to function unambiguously as a causative adverbial. Other forms, such as *as*, *for* and *since*, can have a range of functions, including causative" (Biber 1988: 236). POS trigrams (which are considered to be a light surrogate of syntactic information) have strong discriminating power, but they are corpus dependent (see Argamon et al 1998 and Santini 2004 for POS trigrams extraction methods) and they do not help in the qualitative analysis of a text.

We suggest that the creation and detection of syntactic patterns from a text can represent a useful alternative to the previous approaches to syntax extraction.

4.2.1 Methodology: Creation or Detection of Syntactic Patterns

Syntactic pattern creation and detection involves the following steps:

1. Copying examples of subordinate clauses or sentence types from grammars (namely Werlich 1976; Quirk et al. 1985; Biber et al. 1999) into single files, one file for each syntactic construction.
2. Parsing the files containing the examples of syntactic constructions.
3. Tabulation of the parses in a convenient form, more specifically restoring the horizontal alignment from the Connexor vertical output (see Steps 2 and 3 below).
4. Creation of a set of patterns for each syntactic construction by identifying the common elements of the parses for each syntactic construction and replacing the optional elements with regular expressions.
5. Creation of an algorithm to identify the sets of patterns in running texts.

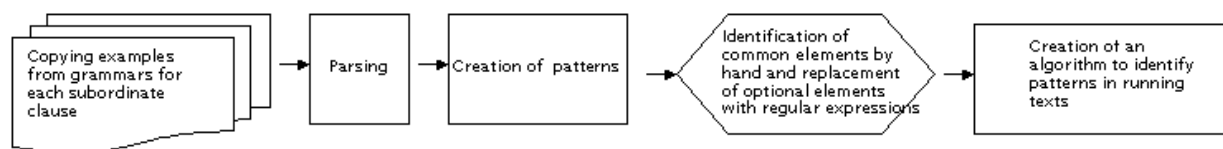


Figure 3. Pipeline for syntactic pattern creation and detection.

All the steps will be illustrated taking the creation and the detection of some patterns of the concessive clause.

Step 1: Grammars as a corpus of examples: Concessive Clause

The examples of clause of concession were copied into a file from Quirk et al. (1985: 1097-1102) and Biber et al. (1999: 818-827).

Step 2: Parsing

The file was parsed in Connexor and a parse was returned for each example (wrong parses were not used to build patterns).

For instance, the sentence:

```
<!-- : Although he had just joined the company, he was treated exactly like
all the other employees.-->
```

was parsed as follows:

1	Although	although	pm:>5	@CS %CS CS
2	he	he	subj:>3	@SUBJ %NH PRON PERS NOM SG3
3	had	have	v-ch:>5	@+FAUXV %AUX V PAST
4	just	just	meta:>5	@ADVL %EH ADV
5	joined	join	cnd:>11	@-FMAINV %VA EN
6	the	the	det:>7	@DN> %>N DET
7	company	company	obj:>5	@OBJ %NH N NOM
8	,	,		
9	he	he	subj:>10	@SUBJ %NH PRON PERS NOM SG3
10	was	be	v-ch:>11	@+FAUXV %AUX V PAST
11	treated	treat	main:>0	@-FMAINV %VP EN
12	exactly	exactly	man:>11	@ADVL %EH ADV
13	like	like		@ADVL %EH PREP
14	all	all	det:>15	@DN> %>N DET
15	the	the	det:>16	@DN> %>N DET
16	other	other	attr:>17	@DN> %>N DET
17	employees	employee		@NH %NH N NOM PL
18	.	.		
19	<s>	<s>		

Figure 4. A sentence parsed by Connexor.

Step 3: Creating the patterns

As described earlier, Connexor returns several types of annotation. For the creation of syntactic patterns, the priority was given to functional annotation, i.e. the annotation starting with an @ sign, occasionally integrated by other types of annotation, for example the syntactic relation `main:>` that identifies the main verb in the whole sentence, or lexical items, such as `although`, to represent subordinators. A simple algorithm based on string manipulation clears out all unnecessary information and returns the following pattern for the sentence of the example:

```
<!-- : Although he had just joined the company, he was treated exactly like
all the other employees.-->
```

```
although@CS @SUBJ have@FAUXV @ADVL FMAINV_EN @DN> @OBJ , @SUBJ be@FAUXV
main_FMAINV_EN @ADVL @ADVL @DN> @DN> @DN> @NH
```

Reading such a pattern is very easy:

```
although@CS      = 'although' in the role of subordinate conjunction
@SUBJ           = subject of the clause
have@FAUXV      = 'have' as auxiliary
@ADVL           = adverb
FMAINV_EN       = past participle
@DN>            = determiner
@OBJ            = object
```

be@FAUXV = 'be' as auxiliary
main_FMAINV = main verb in the sentence
@ADVL = adverb
@NH = noun head

Step 4: Manual Identification of common elements across patterns

What are the common elements between the two following patterns?

Sentence 1:

```
<!-- : Although he had just joined the company, he was treated exactly like
all the other employees.-->
```

```
although@CS @SUBJ have@FAUXV @ADVL FMAINV_EN @DN> @OBJ , @SUBJ be@FAUXV
main_FMAINV_EN @ADVL @ADVL @DN> @DN> @DN> @NH
```

Sentence 2:

```
<!-- : Although Sam had told the children a bedtime story, June told them one
too. -->
```

```
although@CS @SUBJ have@FAUXV FMAINV_EN @DN> @I-OBJ @DN> @A> @OBJ , @SUBJ
main_FMAINV_PAST @I-OBJ @OBJ @ADVL
```

It's very easy to detect: both start with 'although', both have a subject followed by a verb in the subordinate clause, both have a subject and a main verb in the matrix clause. The common elements are:

```
although@CS @SUBJ FMAINV @SUBJ main_FMAINV
```

In order to make this pattern flexible and open to any number of optional elements occurring between each component, we can simply use regular expressions (provided by many programming languages from *Perl* to *Java*, *Javascript*, *Python*, etc. The use of regular expression is well-explained in Friedl 1997). The use of a 'non-greedy' quantifier (i.e. a quantifier that finds the minimum number of character matching the pattern) makes the algorithm more efficient. In many cases, the metacharacters **?* are enough to meet the need of flexibility. The pattern filled in by regular expressions has the following form:

```
although@CS.*?@SUBJ.*?FMAINV.*?@SUBJ.*?main_FMAINV
```

and matches both examples of sentences containing a concessive clause.

Step 5: Creation of an algorithm to identify patterns in running texts

This procedure was repeated for all the examples representing the subordinate clauses and simple sentence types. For each subclause type and for each simple sentence type a set of patterns was built, and a straightforward algorithm to detect them in running texts was created. Each set of patterns was searched by a subroutine, as in the snippet below .

```

sub concessive_clause
{
undef @initial;
$s_v_main = "SUBJ.*?main_FMAINV";

    @initial =
    (
    "although\@CS.*?SUBJ.*?FMAINV.*?$s_v_main",
    "_though\@CS.*?\@NH , .*?$s_v_main",
    "even though\@CS.*?SUBJ.*?FMAINV.*?$s_v_main",
    "even though\@CS.*?\@FMAINV_EN.*?$s_v_main",
    "even if\@CS.*?SUBJ.*?FMAINV.*?$s_v_main",
    );

    for ($count_features = 0; $count_features<@initial; $count_features++)
    {
        if(//(@initial[$count_features]))
        {
            {
                $counter=$counter + 1;
            }
        }
    }
}

```

Figure 5. Example of a subroutine written in Perl to detect the syntactic patterns of concessive clause.

Why not use machine learning?

In a preliminary experiment, we tried to make a classifier to learn a number of patterns but the results were not very encouraging. Features in the dataset were represented by Connexor functional tags or syntactic relations, and data were represented by frequencies of occurrence.

For example, the complex noun phrase *the industrially advanced countries* with the pattern **DN> AD-A> A> NH <s>** was represented by a record with following format:

A>	+FMAINV	-FAUXV	-FMAINV	<AD-A	<NOM	DN>	<P	NH	etc.
1	0	0	0	1	0	1	0	1	

Figure 6. Snippet of a record representing a syntactic pattern.

12 syntactic constructions were selected for the learning task:

- | | |
|--------------------------------|--------------------------------|
| Modifier + Noun Phrase | V + inf-clause |
| Verb + that-clause | V + ing-clause |
| ADjective + that-clause | ADjective + ing-clause |
| that-omission | Linking Adverbials |
| that-retention | Circumstance Adverbials |
| Verb + wh-clause | Existential There |

The training corpus comprised 108 labelled sentences and the test corpus 12 sentences, one per category. The learning situation was extremely favourable, nonetheless, using syntactic patterns made of functional tags, the results were the following:

Classifier	Tags	Training File	Test file
Decision Table	Functional tags	Correctly classified	Correctly classified
		28.70%	25.00%
J48 (C4.5 decision tree)	Functional tags	Correctly classified	Correctly classified
		68.52%	41.67%
IBk (K-nearest neighbors)	Functional tags	Correctly classified	Correctly classified
		98.15%	16.67%
Naive Bayes	Functional tags	Correctly classified	Correctly classified
		69.44%	25.00%

Figure 7. Classification accuracy with functional tags.

Better results were returned when patterns were build with syntactic relations:

Classifier	Tags	Training File	Test file
Decision Table	Syntactic relations	Correctly classified	Correctly classified
		42.26%	33.33%
J48 (C4.5 decision tree)	Syntactic relations	Correctly classified	Correctly classified
		76.19%	75.00%
IBk (K-nearest neighbors)	Syntactic relations	Correctly classified	Correctly classified
		99.07%	25.00%
Naive Bayes	Syntactic relations	Correctly classified	Correctly classified
		78.70%	41.67%

Figure 8. Classification accuracy with syntactic relations.

Syntactic patterns made of syntactic relations can achieve an accuracy of 75% with a decision tree classifier. But as noted earlier, the proportion of the training set makes the situation extremely favourable. It appears that a machine learning approach is not advisable if the aim is only to work out the frequencies of subordinate clauses or other sentence types. First, it is not straightforward to represent the linearity of a string in a classifier. The position of a tag or a word is very important in syntax. The experiment above does not take into account the position of the tag, only the presence and the frequency of it in sentence. The position could be incorporated in the vector, but this would lead to a huge feature space. Of course, it would be possible (see Mitchell 1997: 180 ff.) but the preparation of the vector would be quite complex. Furthermore, evaluating a threshold for the reliability of the predictions returned by the classifier would be an additional problem. Also, multiple classifications of a single sentence (a sentence can include several subordinate clauses) represent another hard tasks to tackle with machine learning.

Advantages of syntactic patterns

The algorithms for creating syntactic patterns from a parser output and their detection in running texts are very easy, straightforward and fast. A syntactic pattern can return lots of syntactic information about main clauses, complement clauses and adverbial

clauses. Also unusual constructions, like the following, can be easily identified, without creating any noise:

Naked as I was, I braved the storm.

```
@PCOMPL-S as@CS @SUBJ be_FMAINV , @SUBJ main_FMAINV @DN> @OBJ
```

In this pattern, there is a subject complement, followed by *as* in the role of conjunction, followed by a subject, followed by a verb *be*, followed by another subject, followed by a main verb, and so on.

With syntactic patterns, we can also detect the position of the subclause in a sentence. The position of a subclause in a sentence can provide hints about registers/genres (see subsections below).

Some subordinators can be easily disambiguated, for example *though*:

No goals were scored, though it was an exciting game.

```
@DN> @SUBJ @FAUXV main_FMAINV , though@CS @SUBJ FMAINV @DN> @A> @PCOMPL-S
```

It cannot be confused with:

He went, though

even if the parser makes a tagging mistake.

Furthermore, several subordinate clauses can be identified in a single sentence. For example:

Although Sam had told the children a bedtime story, June told them one too, **because** she was eager to see their reaction.

both patterns:

```
although@CS @SUBJ.*?FMAINV.*? , @SUBJ.*?main_FMAINV
```

and

```
SUBJ.*?FMAINV.*?because@CS.*?SUBJ.*?FMAINV
```

are applicable, therefore the sentence is fully analysed as bearing a concessive clause in initial position and a reason clause in final position.

4.3 Adverbial clauses

Syntactic patterns for the following subclauses were created. According to Biber et al. (1999: 831), final position is the unmarked choice for non-finite adverbial clauses in all registers. All the syntactic forms of non-finite clauses and all semantic categories typically occur in final position. It is interesting to see which genres or text types, instead, have preferences for the initial position.

- | |
|---|
| <ol style="list-style-type: none">1. concession clause (initial, final, special)2. conditional clause (initial, final, special)3. contrast clause4. exception clause5. purpose clause |
|---|

- | |
|--|
| <ol style="list-style-type: none">6. reason clause (initial, final)7. result clause8. similarity manner comparison clause9. space clause (initial, final)10. time clause (initial, final, incidental, instruct.) |
|--|

Figure 9. List of adverbial clauses.

Legend for the syntactic patterns:

CS=subordinating conjunction;

SUBJ=subject of the clause;

FMAINV=a verb in a clause;

FMAINV_EN=past participle;

FAUX=auxiliary;

main=the main verb in the sentence. It is located in the matrix clause

\$s_v_main=**SUBJ.*?main_FMAINV** (i.e., a subject followed by the main verb of the sentence);

^=beginning of the string;

***?**=non greedy quantifier, it stops at the first occurrence of the following string. It is useful to override all the optional elements in a clause, such as adverbs, adjectives, and so on.

Concession clauses (initial, final, special¹) indicate that the situation in the main clause is contrary to expectations expressed in the concessive clause (Quirk et al. 1097-1098). Clauses of concession are introduced chiefly by *although* or its more informal variant *though*. Other subordinators used with concessive clauses are: *if*, *even if*, *even though*, *when*, *whereas* (formal), *while*, and *whilst*. There are also special constructions, i.e. unusual syntactic ordering when the subordinators are *as* and *though* in formal style, as in *Genius though she was, she was quite unassuming*. Concession may also be expressed by several prepositional phrases followed by a relative clause (*despite*, *in spite of*, *irrespective of*, *regardless of*, *notwithstanding*). They are considered to be stylistically clumsy (Quirk et al. 1985: 1098). The position of the subordinate clause and punctuation marks can help disambiguate ambiguous subordinators. For example, ambiguity between concessive *when* and temporal *when* can be lowered with the help of a comma: concessive *when* is always in final position, often after a comma: *She paid, when she could have entered free (*When she could have entered free, she paid)*, while temporal *when* is more integrated with the main clause: *She was shocked when she heard the story (*She was shocked, when she heard the story; When she heard the story, she was shocked)*. Concessive clauses are widely used in argumentation and argumentative genres (Werlich 1976: 260).

CONCESSIVE CLAUSE, INITIAL POSITION

although@CS.*?SUBJ.*?FMAINV.*?\$s_v_main

Ex: *Although he had just joined the company, he was treated exactly like all the other employees*

though@CS.*?@NH , .*?\$s_v_main

Ex: *Though well over eighty, he can walk faster than I can.*

even though@CS.*?SUBJ.*?FMAINV.*?\$s_v_main

Ex: *Even though you dislike ancient monuments, Warwick Castle is worth a visit.*

¹ The labels "initial", "final", "special" indicate respectively initial position in a sentence, final position in a sentence, and special or unusual syntactic construction of the subclause.

even though@CS.*?@FMAINV_EN.*?\$s_v_main

Ex: *Even though given every opportunity, they would not cooperate with us.*

even if@CS.*?SUBJ.*?FMAINV.*?\$s_v_main

Ex: *Even if you dislike ancient monuments, Warwick Caste is worth a visit.*

whereas@CS.*?SUBJ.*?FMAINV.*? , .*?\$s_v_main

Ex: *Whereas the amendment is enthusiastically supported by a large majority in the Senate, its fate is doubtful in the House.*

The initial position of *whereas* disambiguates the concessive clause from the contrast clause. Usually in the contrast clause, *whereas* is in final position: *I ignore them, whereas my husband is worried of what they think of us.*

CONCESSIVE CLAUSE, FINAL POSITION

\$s_v_main.*?although@CS.*?SUBJ.*?FMAINV

Ex: *He was treated exactly like all the other employees, although he had just joined the company.*

\$s_v_main.*?_though@CS.*?SUBJ.*?FMAINV

Ex: *He can walk faster than I can, though he is over eighty.*

\$s_v_main.*?even though@CS.*?SUBJ.*?FMAINV

Ex: *they would not cooperate with us, even though given every opportunity.*

\$s_v_main.*?even if@CS.*?SUBJ.*?FMAINV

Ex: *Warwick Caste is worth a visit, even if you dislike ancient monuments.*

CONCESSIVE CLAUSE, SPECIAL CONSTRUCTIONS

@NH _though@CS.*?SUBJ.*?FMAINV.*? , .*?\$s_v_main

Ex: *Genius though she was, she was quite unassuming*

main_FMAINV_IMP _though@CS @SUBJ FMAINV",

Ex: *Fail though I did, I would not abandon my goal*

@PCOMPL-S_A _as@CS @SUBJ FMAINV.*? , .*?\$s_v_main

Ex: *Naked as I was, I braved the storm*

Conditional clauses (initial, final, special) mainly express a condition. The two simple subordinators for conditional clauses are *if* and *unless* (Quirk et al. 1985: 1088 ff.). Other subordinators are: *as long as*, *so long as*, *assuming that*, *given that* (formal), *in case*, *in the event that*, *just so that* (informal), *on condition that*, *provided that*, *providing that*, *supposing that*. According to Biber et al. (1999: 833), conditional clauses have different positional distributions across registers, and in written registers, there is a slight preference for the initial position. Special conditional constructions are represented by verbless clauses with *with* or *without*, as in *Without me to supplement your income, you wouldn't be able to manage*, or *With them on our side, we are secure*. Another formal construction is the one with subject inversion : *Had Mark been in charge, it wouldn't have happened*. All these patterns can be easily extracted in an automatic way. *If* can be both concessive and conditional. As we did not manage to disambiguate it for the time being, we assume that it is always conditional, also because concessive *if* seems to be in minority (*If he is poor, he is honest*). Conditional clauses are widely

used in instruction and instructional genres (Werlich 1976: 260), but also in argumentation.

CONDITIONAL CLAUSE, INITIAL POSITION

unless@CS.?SUBJ.*?FMAINV.*?\$s_v_main*

Ex: Unless the strike has been called off, there will be no trains tomorrow.

unless@CS.?FMAINV_EN , .*?\$s_v_main*

Ex: Unless otherwise instructed, you should leave by the back office.

if@CS.?SUBJ.*?FMAINV.*?\$s_v_main*

Ex: If you put the baby down, she'll scream.

if@CS.?FMAINV_EN , .*?\$s_v_main*

Ex: If broken, the pipe won't give you a good smoke.

in_cla_@ADVL_PREP case@<P @SUBJ.?FMAINV.*?@SUBJ.*? FMAINV*

Ex: In case you want me, I'll be in my office till lunch time.

FMAINV_EN _that@CS?SUBJ.*?FMAINV.*?\$s_v_main*

Ex: Provided that you give a month's notice, you might leave the apartment at any time.

CONDITIONAL CLAUSE, FINAL POSITION

\$s_v_main.?unless@CS.*?SUBJ.*?FMAINV*

Ex: There will be no trains tomorrow, unless the strike has been called off.

\$s_v_main.?unless@CS.*?FMAINV_EN*

Ex: you should leave by the back office, unless otherwise instructed.

\$s_v_main.?unless@CS @PCOMPL*

Ex: It has little taste, unless hot.

\$s_v_main.?if@CS FMAINV_EN*

Ex: The grass will grow more quickly if watered regularly.

\$s_v_main.?if@CS @PCOMPL*

Ex: Mary wants me to type a letter if possible.

\$s_v_main.?if@CS @ADVL*

Ex: I can discuss the matter with you now, if necessary.

SUBJ.? FMAINV.*?in_@ADVL_PREP case@<P @SUBJ.*?FMAINV*

I'll be in my office till lunch time, in case you want me.

\$s_v_main.?FMAINV_EN _that@CS*?SUBJ.*?FMAINV*

Ex: You might leave the apartment at any time, provided that you give a month's notice.

\$s_v_main.?FMAINV_ING _that@CS*?SUBJ.*?FMAINV*

Ex: We can leave now, assuming that the movie starts at eight.

\$s_v_main.?just so@ADVL.*?SUBJ.*?FMAINV*

Ex: He does not mind inconveniencing others just so he's comfortable.

"SUBJ.?FMAINV.*?as@AD-A> long@ADVL as@CS.*?SUBJ.*?FMAINV",*

Ex: She may go, as long as he goes with her.

CONDITIONAL CLAUSE, SPECIAL CONSTRUCTIONS

without@ADVL @<P.*?, .*?\$s_v_main

Ex: Without me to supplement your income, you wouldn't be able to manage.

with@ADVL @<P.*?, .*?\$s_v_main

Ex: With them on our side, we are secure.

FAUXV.*?SUBJ.*?FMAINV_EN.*?\$s_v_main

Ex: Had Mark been in charge, it wouldn't have happened.

Contrast clauses are introduced by several subordinators that introduce also concessive clauses (Quirk et al. 1985: 1102): *whereas*, *while* (also temporal) and *whilst*. The contrastive meaning may be emphasized by *in contrast* and *by contrast* when the contrastive clause is initial. While concessive *while* is always in initial position, as in *While he has many friends, Peter is lonely*, temporal *while* can be in both positions: initial position, as in *While I was asleep, I dreamt of you*; or final position without a comma, as in *They arrived while I was sunbathing*. Contrast clauses controlled by *while* in final position are very unambiguous in this respect, because they always follow a comma, as in: *John teaches physics, while Mary teaches chemistry* (**They arrived, while I was sunbathing*).

\$s_v_main.*?, while.*?SUBJ.*?FMAINV

Ex: Mr Larson teaches physics, while Mr Corby teaches chemistry.

cf. the concessive pattern for 'whereas' above.

\$s_v_main.*?, whereas.*?SUBJ.*?FMAINV

Ex: I ignore them, whereas my husband is always worried about what they think of us.

Exception clauses are introduced by several subordinators: *but that* (formal), *except that*; less frequently *excepting that* and *save that* (formal) (Quirk et al. 1985: 1102). They seem to occur only in final position.

\$s_v_main.*?except@CS.*?SUBJ.*?FMAINV

Ex: I would pay you now, except I don't have money on me.

SUBJ.*?FMAINV.*?save@FMAINV _that@CS.*?SUBJ.*?FMAINV

Ex: No memorial remains for the brave who fell on that battlefield, save that they will leave their image for ever in the hearts and minds of their grateful countrymen.

SUBJ.*?FMAINV.*?but@CC _that@CS.*?SUBJ.*?FMAINV

Ex: Nothing would satisfy the child but that I place her on my lap.

SUBJ.*?FMAINV.*?but for@ADVL @<P @INFMARK>

Ex: Nothing would satisfy the child but for me to place her on my lap.

SUBJ.*?never.*?FMAINV.*?but.*@CC.*?SUBJ.*?FMAINV

Ex: It never rains but it pours.

Purpose clauses are more often infinitival than finite. More explicit subordinators of purpose are *in order to* (formal) and *so as to*. Finite clauses of purpose are introduced by *so that* or (less commonly and more informally) by *so*, and (more formally) by *in order that* (Quirk et al. 1985: 1107). Purpose clause are syntactically very similar to infinitival nominal clause (*I wished to go home vs. I left early to catch the train*). A way of disambiguating them at a syntagmatic level is to use lexicalized verbs, because all complement/nominal clauses are controlled by restricted sets of verbs (Biber et al 1999:

700 ff.). Purpose clauses are widely used in instruction and instructional genres (Werlich 1976: 266).

PURPOSE CLAUSE, INITIAL POSITION

`^@INFMARK>.*?FMAINV_INF.*?main)`

Ex 1: To be neutral in this conflict is out of the question

Ex 2: To open the carton, pull this tab

`main_FMAINV.*?so@ADVL _that@CS.*?SUBJ.*?FAUX.*?FMAINV`

Ex: The school closes earlier so that the children can get home before dark.

PURPOSE CLAUSE, FINAL POSITION

`^(.*?)main_FMAINV(.*)INFMARK>.*? FMAINV_INF`

The first group of parentheses is used to check that the main verb of the sentence is not one of the verbs supporting infinitival “to”, for example *want*, *wish*, and so on (the full list has been taken from Biber et al. (1999: 700-705) and Quirk et al. (1985: 1181 ff.). The algorithm checks that none of these verbs is in the part of the string between brackets. If so, the pattern is not matched.

The second group of parentheses is used to check that there are no subordinators between the main verb of the sentence and the infinitival tag (**INFMARK**). The algorithm checks that none of the subordinators used in for the syntactic pattern creation is in the part of the string between brackets. If so, the pattern is not matched.

PURPOSE CLAUSE, ANY POSITION

`in@ADVL order@<P for@ADVL @<P @INFMARK>.*?FMAINV_INF`

Ex: They left the door open in order for me to hear the baby.

`in order to@INFMARK>.*?FMAINV_INF`

Ex: The committee agreed to adjourn in order to reconsider the matter when fuller information became available.

`in order that@CS.*?FMAINV`

Ex: The jury and the witnesses were removed from the court in order that not hear the arguments.

`so as to@INFMARK>.*?FMAINV_INF`

Ex: Students should take notes so as to make revision easier.

Reason clauses (initial, final) include several types of subordinate clauses. For all types, the word 'reason' is a superordinate term (cause and effect; reason and consequence; motivation and result; circumstances and consequence) (Quirk et al. 1985: 1103 ff.). Reason clauses are most commonly introduced by the subordinators *because* and *since*. Other subordinators are *as* and *for* (formal). *Because* is unambiguous, *since* can be also temporal. However, temporal *since* comes always with present perfect in the matrix clause, while reason *since* does not (see the example below). *For* is disambiguated by the tag **@CS**, when it functions as subordinator. *As* (used for reason, similarity and time) cannot be disambiguated yet. Reason clauses in final position are the strong preference for conversational genres (Biber et al. 1999: 833). Reason clauses are widely used in exposition (Werlich 1976: 254) and argumentation (Werlich 1976: 260).

REASON CLAUSE, INITIAL POSITION

`because@CS.*?SUBJ.*?FMAINV.*?$s_v_main`

Ex: Because we live near the sea, we often go sailing. (inventato)

`since@CS.*?SUBJ.*?FMAINV.*?SUBJ(.*)FMAINV",`

Ex: Since we live near the sea, we often go sailing.

The group of parentheses is used to check that the verb in the matrix clause is not a present perfect or a past perfect. These forms are allowed in the matrix clause only when "since" is temporal, like in: *Since they lived in London, they have been increasingly happy* (Quirk et al. 1985: 538 ff.), and not when it is causal.

REASON CLAUSE, FINAL POSITION

`$s_v_main.*?because@CS.*?SUBJ.*?FMAINV`

Ex: I lent him the money because he needed it.

`$s_v_main.*?for@CS.*?SUBJ.*?FMAINV`

Ex: Much has been written about psychic phenomena, for they pose fascinating problems that have yet to be resolved.

`SUBJ.*?FMAINV.*?since@CS.*?SUBJ(.*)FMAINV`

Ex: I lent him the money because he needed it.

Result clauses are introduced by the subordinators *so that* (formal), and *so*. These clauses overlap with those of purpose both in meaning and in subordinators. The main semantic difference is that result clauses are factual and not putative, i.e. in result clauses the result is achieved, whereas in purpose clauses it is a desired result (Quirk et al. 1985: 1108 ff.). Therefore, clauses of result do not require a modal auxiliary (*We paid him immediately, so he left contented*), while purpose clauses often have one (*We paid him immediately so he would leave contented*). Result clauses can only appear in final position. Unlike purpose clauses, result clauses introduced by *so* are separated by a comma (Quirk et al. 1985: 1109). Result clauses are widely used in description and descriptive genres (Werlich 1976: 254).

`$s_v_main.*? , _so@ADVL _that@CS.*?SUBJ(.*)FMAINV`

Ex: We paid him immediately, so that he left content.

`$s_v_main.*? , _so@CS.*?SUBJ(.*)FMAINV`

Ex: We paid him immediately, so he left content.

The group of parentheses is used to check that the verb in the matrix clause is not a preceded by an auxiliary.

Similarity, manner and comparison clauses are introduced by *as* and *like* (similarity and manner) and *like* (informal), *as if*, *as though* (comparison combined with manner) (Quirk et al. 1985: 1110-1111). Manner clauses are widely used in description and instruction (Werlich 1976: 267).

`as@CS.*?SUBJ.*?FMAINV.*? , _so@CS.*?$s_v_main`

Ex: As a moth is attracted by a light, so he was attracted by her.

`$s_v_main.*?as if@CS.*?SUBJ.*?FMAINV`
Ex: He looks as if he's getting better.

`$s_v_main.*?as if@CS @FMAINV_ING`
Ex: He bent down as if tightening his shoe laces.

`$s_v_main.*?as if@CS @INFMARK> @FMAINV that@CS`
Ex: She winked at me as if to say that I shouldn't say anything.

`$s_v_main.*?as though@CS.*?SUBJ.*?FMAINV`
Ex: She treated me as though I was a stranger.

Space clauses (initial, final) are introduced mainly by *where* or *wherever* (Quirk et al. 1985: 1087). They indicate position or direction. Space clauses are widely used in description (Werlich 1976: 254).

SPACE CLAUSE, INITIAL POSITION

`where.*?SUBJ.*?FMAINV.*?$s_v_main`
Ex 1: Where I saw only wilderness, they saw abundant signs of life.
Ex 2: Wherever I saw only wilderness, they saw abundant signs of life.

`where.*?SUBJ.*?FMAINV.*?,.*?$s_v_main`
Ex: Where the fire had been, we saw nothing but blackened ruins.

SPACE CLAUSE, FINAL POSITION

`($s_v_main.*?)where.*?SUBJ.*?FMAINV`
Ex 1: They went where they could find work.
Ex 2: They went wherever they could find work.

The group of parentheses is used to check that the main verb of the sentence is not one of the verbs supporting wh-clauses, as in the example: *I want you to show me where the car went down*. Otherwise, the pattern is not matched.

Time clauses (initial, final, incidental, instructional) are introduced by one of the following subordinators: *after, as, before, once, since, till, until, when whenever, while, whilst, now that, as long as, so long as, as soon as, immediately, directly*. Adverbial -ing clauses of time are introduced by one of the following subordinators: *once, till, until, when, whenever, while, whilst*. Adverbial -ed clauses of time are introduced by one of the following subordinators that are also used with finite clauses: *as soon as, once, till, until, when, whenever, whilst*. Verbless clauses of time are introduced by the same subordinators as -ed clauses. *Until* and *till* in verbless clause are mainly used in instructional language (Quirk et al. 1985: 1079). -ing clauses without a subordinator or a subject may also express time relationship, as in *Returning to my village after 15 years, I met an old schoolteacher* (see Quirk et al. 1985: 1078 ff.). Time clauses are widely used in narration (Werlich 1976: 256) and in instruction (Werlich 1976: 267).

TIME CLAUSE, INITIAL POSITION

`since@CS.*?SUBJ.*?FMAINV.*?SUBJ.*?have@FAUX.*?FMAINV`
Ex: Since I saw her last, she has dyed her hair.

`once.*?FMAINV.*? , .*?$s_v_main`
Ex: Once having made a promise, you should keep it.

when.*?SUBJ.*?FMAINV.*?\$s_v_main
Ex: *When I last saw you, you lived in Washington.*

when.*?FMAINV_ING.*? , .*?FMAINV
Ex: *when crossing the street be careful.*

Initial “when” is always temporal. Concessive “when” is always in final position, as in:
She paid when she could have entered free (Quirk et al. 1985: 1097).

TIME CLAUSE, FINAL POSITION

FMAINV.*? as@AD-A> soon_tmp_@ADVL_ADV as@CS.*?SUBJ.*?FMAINV
Ex: *Buy your ticket as soon as you reach the station.*

FMAINV.*? as@AD-A> soon_tmp_@ADVL_ADV as@CS FMAINV_EN
Ex: *The document will be returned as soon as completed.*

FMAINV.*?as@AD-A> soon_tmp_@ADVL_ADV as@ADVL @<P
Ex: *Complete your work as soon as possible.*

FMAINV.*? , never@ADVL @INFMARK> FMAINV_INF
Ex: *He left, never to return.*

SUBJ.*?have@FAUX.*?FMAINV.*?since@CS.*?SUBJ.*?FMAINV
Ex: *She has dyed her hair, since I saw her last.*

SUBJ.*?FMAINV.*?immediately.*?SUBJ.*?@FAUXV @FMAINV_EN
Ex: *I'll give an answer immediately I've finished reading your file.*

\$s_v_main.*?before.*?FMAINV_ING
Ex: *They washed their hands before eating.*

\$s_v_main.*?after.*?FMAINV_ING
Ex: *They washed their hands after eating.*

on.*?@<PFMAINV_ING.*? , .*?\$s_v_main
Ex: *On becoming a member, you will receive a membership card and a badge.*

\$s_v_main.*?when.*?@FMAINV_EN
Ex: *Spinach is delicious when eaten raw.*

\$s_v_main.*?while@ADVL @FMAINV_EN
Ex: *He slept while stretched out on the floor.*

\$s_v_main.*?until@CS.*?FMAINV_EN
Ex: *The dog stayed at the entrance until told to come it.*

FMAINV.*?when.*?@FMAINV_ING
Ex: *Be careful when crossing the streets;*

TIME CLAUSE, INCIDENTAL POSITION

SUBJ.*? , once.*?SUBJ.*?FMAINV.*? , .*?main_FMAINV
Ex: *My family, once they saw the mood I was in, left me completely alone.*

SUBJ.*? , when.*?.*?SUBJ.*?FMAINV.*? , .*?main_FMAINV
Ex: *All applications, when they are received before the deadline, are dealt with promptly.*

TIME CLAUSE, INSTRUCTIONAL

SUBJ.*?FAUX.*?FMAINV.*?until@ADVL @<P
Ex: You should beat the mixture until fluffy.

main_FMAINV_IMP.*?until@ADVL @<P
Ex: Beat the mixture until fluffy.

when@ADVL.*?@<P , .*?@FMAINV_IMP
Ex: When in difficulty, consult the manual.

when@ADVL.*?@<P , .*?SUBJ.*?FAUX.*?FMAINV
Ex: When in difficulty, you should consult the manual.

4.4 Complement/Nominal Clauses

Complement clauses are a type of dependent clause used to complete the meaning of an associated verb or adjective in a higher clause in the sentence. Complement clauses are also called nominal clauses because they occupy a noun phrase slot. There are four major structural types of complement clause: *that*-clause, *wh*-clause, *to*-infinitive clause, *ing*-clause (Biber et al. 1999: 658). Syntactic patterns for the following nominal clauses were created, mainly on post-predicate position:

1. verb+*that* clause

2. adjective+*that* clause

3. *that* omission

4. *wh*-clause

5. verb+*to* clause

6. adjective+*to* clause

7. verb+*ing* clause

8. comparative clause

9. relative clause

Figure 10. List of nominal clauses.

In general, *that*-clauses in post-predicate position are commonly used to report speech, thoughts, attitudes, or emotions of humans (Biber et al. 1999: 661). They are also widely used in argumentation (Werlich 1976: 261).

Verb+*that* clause. Verbs taking *that* complement clauses in post-predicate position fall into three semantic domains: mental verbs, speech act verbs, and other communication verbs (Biber et al. 1999: 661). Post-predicate *that*-clauses after verbs are most common in conversational genres, and less common in academic genres (Biber et al. 1999: 674).

main_FMAINV.*?_that@CS.*?SUBJ.*?FMAINV

Ex 1: They warned him that it's dangerous.

Ex 2: He was told that she had checked out of the hospital.

Adjective+*that* clause. The adjectives that control a *that* complement clause convey stance, and fall into three major semantic domains: degree of certainty, affective psychological states, and evaluation of situations, events, etc. (Biber et al. 1999: 671). Post-predicate *that*-clauses controlled by adjectival are most common in conversational genres (Biber et al. 1999: 674).

PCOMPL-S_A _that@CS.*?SUBJ.*?FMAINV

Ex: was quite confident that it would stay in very well.

@<P _that@CS.*?\$s_v_main

Ex: It has been clear for some time that the demands of the arms control process would increasingly dominate military planning.

that omission. According to Biber et al. 1999: 680 ff., there are a number of discourse factors that influence the retention vs. omission of *that*. In conversation, the omission is the norm, while in writing is the opposite. These distributional patterns correspond to different production circumstances and communicative purposes.

SUBJ.*?main(.*)SUBJ.*?FMAINV

Ex: I hope you realised they said a few words on there.

be_main.*?@PCOMPL-S_A(.*)SUBJ.*?FMAINV

Ex: I'm sure he was wrong.

be_main.*?@PCOMPL-S_NOM(.*)SUBJ.*?FMAINV

Ex: It's a pity you don't know Russian.

FMAINV.*?FMAINV_EN(.*)SUBJ.*?FMAINV

Ex: I'll do it provided you pay me.

be@FAUXV.*?FMAINV_EN(.*)SUBJ.*?FMAINV",

Ex: It was hoped by everybody she would sing.

The group of parentheses in the patterns indicate that the part of strings between brackets should not contain "that" or any other subordinators. If so, the pattern is not matched.

wh-clause. Wh-clauses can be either dependent interrogative clauses or nominal relative clauses. Dependent interrogative clauses are used with verbs such as *ask* or *wonder* to ask an indirect question, as in *I wonder what this is about*. In contrast, nominal relative clauses have a more complicated structure, as in *Whoever sorts Eagle out will make a lot of money* (Biber et al. 1999: 683 ff.). Wh-clauses are more common in conversational genres (Biber et al. 1999: 688 ff.).

WH CLAUSE, INITIAL POSITION

which@SUBJ be_main where_loc_@ADVL_WH.*?SUBJ.*?FMAINV

Ex: Which is where Carlos used to live.

which@SUBJ be_main when_@ADVL_WH.*?SUBJ.*?FMAINV

Ex: Which is when Carlos used to come.

what@SUBJ FMAINV.*?be_main how@ADVL.*?SUBJ.*?FMAINV

Ex: What baffles me is how few of them can spell.

what@SUBJ be_FMAINV.*?PCOMPL-S.*?be_main.*?PCOMPL-S

Ex: What is good among one people is an abomination with others.

what@DN>.*?SUBJ.*?FMAINV.*?_main

Ex: What a single mother represents may seem touchingly attractive.

what@PCOMPL-S.*?_main.*?FMAINV

Ex: What could be at work there is an actual enmity towards the very structure of society.

what@SUBJ.*?FMAINV.*?_main

Ex: What to came to is this.

how@ADVL @INFMARK> @FMAINV.*?@SUBJ.*?@FMAINV

Ex: *How to read the record is the subject of much of this book.*

whoever@SUBJ FMAINV.*?FMAINV

Ex: *Whoever sorts Eagle out will make a lot of money.*

who@SUBJ FMAINV.*?FMAINV

Ex: *The person who sorts Eagle out will make a lot of money.*

WH CLAUSE, FINAL POSITION

that@SUBJ.*?_main what@OBJ.*?SUBJ.*?FMAINV

Ex: *That's what the case is all about.*

that@SUBJ.*?_main why@ADVL.*?SUBJ.*?FMAINV

Ex: *That's why I bought the refill.*

main_FMAINV.*?what.*?SUBJ.*?FMAINV

Ex: *I don't know what they are.*

main_FMAINV.*?which.*?SUBJ.*?FMAINV

Ex: *I didn't know which one you liked best.*

main_FMAINV.*?how.*?SUBJ.*?FMAINV

Ex: *I can remember how I used to be.*

main_FMAINV.*?where.*?@SUBJ.*?FMAINV

Ex: *I want you to show me where the car went off.*

be_main.*?PCOMPL-S_A where.*?SUBJ.*?FMAINV

Ex: *I'm not sure where the meeting can be arranged.*

main_FMAINV.*?when.*?@SUBJ.*?FMAINV

Ex: *I want you to tell me when this happened.*

be_main.*?PCOMPL-S_A when.*?SUBJ.*?FMAINV

Ex: *I'm not sure when it's open.*

Verb+to clause. The verbs taking *to*-clauses in post-predicate position can be grouped into a number of semantic classes (Biber et al. 1999: 693 ff.). According to Biber et al.(1999: 711), *to*-clauses in post-predicate position typically perform specific functions in different registers.

FMAINV.*?FMAINV.*?@INFMARK>.*?FMAINV

Ex: *I'm just trying to get away early.*

I'm just trying to get away early.

Adjective+to clause. Adjectival predicates controlling *to*-clauses come from a number of semantic domains (Biber et al. 1999: 718 ff.). Post-predicate *to*-clauses complementing an adjective are rare in conversational genres and common in press genres (Biber et al. 1999: 722).

PCOMPL-S_A.*?@INFMARK>

Ex: *Millar was obstinately determined to change the content of education.*

Verb+ing clauses are most commonly used in conjunction with an aspectual verb in the main clause (*begin, start, stop*), but they are also used to report speech acts, cognitive states, perceptions, emotions, and various other actions. They are mostly located in

post-predicate position (Biber et al. 1999: 739). Overall, -ing clauses in post-predicate position are most common in the writing and least common in conversational genres (Biber et al. 1998: 749).

FMAINV(.*) FMAINV_ING

Ex 1: I remember reading this book.

Ex 2: They talk about building more.

Comparative clause. In a comparative construction, a proposition expressed in the main clause is compared with the proposition expressed in the subordinate clause (Quirk et al. 1985: 1127 ff.). The frequency of comparative clauses in academic writing reflects the importance of comparison as a means of understanding and explicating reality (Biber et al. 1999: 728-729.). Comparative clauses are also widely used in description and descriptive genres (Werlich 1976: 254).

@SUBJ.*?main.*?@PCOMPL-S_CMP than@ADVL

Ex: Jane is healthier than her sister.

@SUBJ.*?main.*?@PCOMPL-S_A than@ADVL

Ex: Mary is less old than Jane.

@SUBJ.*?main.*?@OBJ @ADVL_CMP

Ex: James enjoys the theatre more.

@SUBJ.*?main.*?@AD-A_CMP @ADVL

Ex: Time passed more quickly than last year.

@SUBJ.*?main.*?_as@AD-A> @SUBJ _as@CS @SUBJ

Ex: I agree with you as much as I can.

@SUBJ.*?main.*?_as@AD-A> @PCOMPL-S_A _as@CS

Ex: The article was as objective as I expected.

@SUBJ.*?main.*?@PCOMPL-S_A _enough@<AD-A @INFMARK>

Ex: You are old enough to look after yourself.

@SUBJ.*?main.*?_enough@OBJ.*?@INFMARK>

Ex: She knows enough about the topic to explain it to you.

Relative clauses. Restrictive/nominal relative clauses are different from sentential relative clauses. While restrictive/nominal relative clauses have a noun as antecedent, the sentential relative clause refers back to the predicate or to a whole sentence. For example, *Things then improved, which surprises me* is a sentential relative clause (Quirk et al. 1985: 1118) and *I eat what I like* is a restrictive relative clause (Quirk et al. 1985: 1056). In standard English, relative clauses can be formed using eight different relativisers: *which, who, whom, whose, that, where, when, why*, plus a zero relativiser. The choice among relativisers is influenced by a number of factors, for example register, restrictive vs. non-restrictive function, animate vs. non animate head. Restrictive relative clauses are widely used in description (Werlich 1976: 254) and exposition (Werlich 1976: 258). Relative clauses are currently extracted using two tags:

<Rel>

Ex 1: She knows enough about the topic to explain it to you.

Ex 2: They are delighted with the person who has been appointed.

who@DN>

Ex: *The woman whose daughter you met is Msr. Brown.*

4.5 Simple Sentences

In his text grammar for English, Werlich, among other things, lists six *phenomenon sentences* (Werlich 1976: 216 ff.). According to his cognitive framework, all sentences in the communication process interact with contextual factors of the spatio-temporal environment. Against this referential context, all sentences are phenomenon sentences because the text producer emphasizes a type of phenomenon in the spatio-temporal context. According to Werlich, when the frequency of one of these main clauses is very high in a single text, the text is likely to belong to one of the five text type dealt with in his text grammar. As noted earlier, Werlich's terminology does not belong to the English tradition. According to him, the English language provides the following six types of simple phenomenon sentences:

1. <i>phenomenon registering</i>
2. <i>action recording</i>
3. <i>phenomenon identifying</i>

4. <i>phenomenon linking</i>
5. <i>quality attributing</i>
6. <i>action demanding</i>

Figure 11. Werlich's phenomenon sentences.

The **phenomenon registering** sentence is used with reference to phenomena in space. It is a simple subject-predicate-adverbial structure with an adverb of place. The adverb of place in the phenomenon registering sentence sets a spatial frame of reference within which the phenomena at the subject are placed, as in *Thousands of glasses were on the table*. A variant of the phenomenon registering sentence is the pattern with *there* followed by a verb form of *be*. Phenomenon registering sentences are widespread in description (Werlich 1976: 216-217).

@SUBJ.*?main.*?loc_@ADVL_PREP

Ex: *Thousands of pilgrims were on the road to London.*

@SUBJ.*?main.*?pth_@ADVL_PREP

Ex: *Thousands of tourists moved to London.*

@SUBJ.*?main.*?sou_@ADVL_PREP

Ex: *Thousands of tourists moved away from London.*

@F-SUBJ be_main.*?@PCOMPL-S_NOM

Ex: *There was a long queue of cars.*

The **action recording** sentence is used with reference to phenomena in time. It is a simple subject-predicate-adverbial-(adverbial) structure with adverbs of place and time. The adverb of time in the action recording sentence sets a temporal frame of reference within which the phenomena at subject are placed, as in *The passengers landed in New York in the middle of the night*. Action recording sentences are widespread in narration (Werlich 1976: 217).

SUBJ.*?main.*?tmp_@ADVL

Ex: *He went home for Christmas.*

"SUBJ.*?main.*?frq_@ADVL",
Ex: he went back home often.

SUBJ.*?main.*?dur_@ADVL
Ex: this happened in the first 25 years.

cla_@ADVL_PREP @<P
Ex: At 10, the engine started.

The **phenomenon identifying** sentence is used with reference to phenomena in relation to concepts. It is a simple subject-predicate-complement structure with a verb form of *be* at predicate and a nominal group at complement, as in *One part of the brain is the cortex*. The nominal group at the complement identifies the phenomenon referred to in the nominal group at subject giving it a name, a class or a role. Phenomenon identifying sentences in the present are common in exposition (Werlich 1976: 217-218).

@SUBJ be_main.*?@PCOMPL-S_NOM
Ex: These are nerve cells or neurones.

The **phenomenon linking** sentence is used with reference to phenomena in *part-of* relationship with other phenomena. It is a three element patterns, subject-predicate-complement, with a verb form of *have* at predicate and a nominal group at the complement, as in *The brain has ten million neurones*. Phenomenon-linking sentences in the present tense are common in exposition (Werlich 1976: 218).

SUBJ have_main.*?OBJ
Ex: The brain has a network of swichtes.

The **quality attributing** sentence is used with reference to qualities and properties of phenomena. It is a simple subject-predicate-complement structure with a verb form of *be* at predicate and an adjective (or equivalent) at the complement, as in *The obsession with durability in the arts is not permanent*. Negative quality attributing sentences are common in argumentation (Werlich 1976: 218-219).

SUBJ.*?be_main.*?PCOMPL-S_A
Ex: Endurance in art forms is essential, claims George.

The **action demanding** sentence is used with reference to behavioral responses from the addressee. Its basic form is the command, as in *Be reasonable for a moment!*. Action demanding sentences are common in instructional texts (Werlich 1976: 219).

main_FMAINV_IMP.*?OBJ
Ex: Add a layer of pebbles to prevent the soil being washed away.

@FAUXV_IMP.*?OBJ
Ex: Do not add fertilizer or manure.

5 Preliminary Results: Accuracy of Linguistic Facets for Web Genre Classification

As pointed out in the introduction, linguistic facets are linguistically-motivated and functionally-interpretable features designed to combine qualitative analysis and

quantitative findings to get a more accurate text analysis together with genre and text type identification. In this section we wish to give a hint about their discriminating power.

We selected four web genres -- *blogs, FAQs, front pages, and personal home pages* -- and collected 200 web pages per genres from public specialized web archives. Then we built two SVM classifiers with the Weka package (Witten and Frank 2000). One model was built using 211 POS trigrams and the other one with 84 linguistic facets.

There are several ways of selecting POS trigrams. The total amount of POS trigrams in a corpus can be several thousands. First we computed POS trigrams per each web genre, and deleted those POS trigrams with a frequency lesser than 200 (each web genre was represented by 200 web pages, so in the most optimistic hypothesis a POS trigram was present at least once in each of the web pages). Then we collected all the POS trigrams for all the four web genres into a single file, and sorted them into a list ignoring their frequencies of occurrence in the single web genres. Finally we deleted all the POS trigrams that appeared more than twice in this list. The rationale behind this choice is that in order to be discriminating, a POS trigram must not be present in all web genres, and we decided a threshold of two out of four web genres. We ended up with 211 POS trigrams.

The 84 linguistic facets comprised 27 functional cues, 13 connectives, 29 subordinate patterns, 9 simple sentence patterns (we added 3 variants to the patterns presented in subsection 4.5), and 6 subordinators not fully disambiguated with syntactic patterns (*if, as, while, whereas, when, and since*)

Linguistic facets are mainly based on syntactic annotation, and POS trigrams represent a shallow approach to syntax. Both of them have a syntactic nature so they can be compared consistently. We ran the classifiers ten times with ten different seeds and averaged the accuracy results. 211 POS trigrams returned an averaged accuracy of 86.50%, while linguistic facets returned an averaged accuracy of 84.28%. According to these figures, the performance of linguistic facets is very promising, considering that this is only an initial set of 84 features.

6 References

- Aaronson S. (1999), *Stylometric Clustering: A comparative analysis of data-driven and syntactic features*, Project report available at <http://www.cs.berkeley.edu/~aaronson/sc/report.doc>
- Argamon S., Koppel M., Avneri G. (1998), Routing documents according to style, *Proceedings of the First International Workshop on Innovative Internet Information Systems (IIS-98)*.
- Baayen H., Halteren H. van, Tweedie F. (1996), Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, *Literary and Linguistic Computing*, 11.
- Biber, D. (1988), *Variations across speech and writing*, Cambridge University Press, Cambridge.
- Biber, D. (1989), A typology of English texts, *Linguistics*, Vol. 27, 3-43.
- Biber D., Johansson S., Leech G., Conrad S., Finegan E. (1999), *Longman Grammar of Spoken and Written English*, Longman, Harlow.
- Copeck T., Barker K., Delisle S. and Szpakowicz S. (2000), Automating the Measurement of Linguistic Features to Help Classify Texts as Technical, *Conférence TALN 2000, Lausanne*, 16-18 October 2000.
- Dewdney N., VanEss-Dikema C., MacMillan R. (2001), The form is the Substance: Classification of Genres in Text, *ACL '2001 Conference*, Toulouse, France.
- Friedl, J. (1997), *Mastering Regular Expressions*, O' Reilly, Beijing, Cambridge, etc.
- Karlgren J. (2000), *Stylistic Experiments for Information Retrieval*, Thesis submitted for the degree of Doctor of Philosophy, Department of Linguistics, Stockholm University.
- Kessler B., Numberg G., Shütze H. (1997), Automatic Detection of Text Genre, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*.
- Lim C. S., Lee K. J., Kim G. C. (2005), Automatic Genre Detection of Web Documents, Su K., Tsujii J., Lee J., Kwong O. Y. (eds.) *Natural Language Processing – IJCNLP 2004*, Springer, Berlin Heidelberg.
- Michos S., Stamatatos E., Fakotakis N., Kokkinakis G. (1996), An Empirical Text Categorizing Computational Model Based on Stylistic Aspects, *Proceedings of the 8th International Conference on Tools with Artificial Intelligence (TAI'96)*.
- Mitchell T. (1997), *Machine Learning*, McGraw-Hill International Editions, New York.
- Nakamura, J. (1993), Statistical Methods and Large Corpora – A New Tool for Describing Text Type, in *Text and Technology*, Baker M., Francis G., Tognini-Bonelli E. (eds.), J. Benjamins Publishing Company, Philadelphia - Amsterdam, pp. 291-312.

- Quirk R., Greenbaum S., Leech G., Svartvik J. (1972), *A Grammar of Contemporary English*, Longman.
- Quirk R., Greenbaum S., Leech G., Svartvik J. (1985), *A Comprehensive Grammar of the English Language*, Longman.
- Santini M. (2004a), *State-of-the-art on Automatic Genre Identification*, Technical Report ITRI-04-03, 2004, ITRI, University of Brighton (UK).
- Santini M. (2004b), *A Shallow Approach To Syntactic Feature Extraction For Genre Classification*, *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 04)*, University of Birmingham (UK), 6-7 January, 2004.
- Santini M. (2005), *Annotated corpora vs. raw web page collections. Text types, web pages, and Linguistic features: Some issues*, *AAACL/ICAME*, 12-15 May 2005, Ann Arbor, MI, USA.
- Sigley R. (1997), *Text Categories and Where You can Stick Them: A Crude Formality Index*, *International Journal of Corpus Linguistics*, Vol. 2 No. 2, pp. 199-237.
- Stamatatos E., Fakotakis N., Kokkinakis G. (2000), *Text Genre Detection Using Common Word Frequencies*, *Proceedings of the 18th International Conference on Computational Linguistics (COLING2000)*.
- Tapanainen P., Järvinen T. (1997), *A non-projective dependency parser*, *Proceedings of the 5th Conference on Applied Natural Language Processing*.
- Werlich E. (1976), *A Text Grammar of English*, Quelle & Meyer, Heidelberg (Germany).
- Wikberg, K. (1993), *Verbs as indicators of text type and/or style: Some observations on the LOB corpus*, in *Corpus-based Computational Linguistics*, Souter C., Atwell E. (eds.), Rodopi, Amsterdam-Atlanta, pp. 127-145.
- Witten I. and Frank E. (2000), *Data Mining: Practical machine learning tools with Java implementations*, organ Kaufmann, San Francisco.
- Wolters M., Kirsten M. (1999), *Exploring the Use of Linguistic Features in Domain and Genre Classification*, *Proceedings of EACL '99*, pages 142-149.