Variation in the contextuality of language: an empirical measure

FRANCIS HEYLIGHEN* & JEAN-MARC DEWAELE**

**Center "Leo Apostel", Free University of Brussels, Pleinlaan 2, B-1050 Brussels, Belgium; fheyligh@vub.ac.be; http://pespmc1.vub.ac.be/HEYL.html*

*** Birkbeck College, University of London, 43 Gordon Square, WC1H 0PD London, United Kingdom; j.dewaele@french.bbk.ac.uk*

Abstract: The concept of formality/contextuality is proposed as the most important dimension of variation between linguistic expressions. Formal communication conveys information explicitly, through the linguistic expression itself, whereas contextual communication conveys information implicitly, through the context of the expression. An empirical measure of formality, the F-score, is proposed, based on the frequencies of different word classes. Nouns, adjectives, articles and prepositions are more frequent in formal styles; pronouns, adverbs, verbs and interjections are more frequent in contextual styles. This measure adequately distinguishes different genres of language production using data for Dutch, French, Italian, and English. Factor analyses applied to data in 7 different languages produce a similar factor as the most important one. Both the data and the theoretical model suggest that contextuality decreases when unambiguous understanding becomes more important or more difficult to achieve, when the separation in space, time or background between the interlocutors increases, and when the speaker is male, introverted and/or academically educated.

Keywords: contextuality, formality, language, word frequencies, personality, situation.
Submitted to: Foundations of Science

1. Introduction

One of the fundamental issues when studying context is to determine the degree of context-dependence in a given situation. All communication or linguistic expression necessarily refers to the context to some degree (Heylighen, 1999), but in some situations context will obviously play a much higher role than in others.

The anthropologist Edward T. Hall (1976) has distinguished two fundamental types of situations: *high-context* and *low-context*. In low-context situations, communication is explicit and overt, stating the facts exactly and in detail. In high-context situations, communication is implicit, and information is conveyed more by the context than by the verbal expression. Although Hall introduced this concept primarily to distinguish different types of cultures (e.g. American and Northern European cultures are typically low-context, while Mediterranean and Eastern cultures are high-context), the same distinction can be applied to different communicative situations within the same culture. For example, twins who have grown up together will be able to make themselves understood with a minimum of explicit communication (high-context), while lawyers in a courtroom need to formally state all their assumptions, arguments and inferences (low-context) (Hall, 1976).

Such distinctions between high and low context situations in cultural anthropology are largely based on personal experience and on global impressions of how people in a particular culture behave. Moreover, the association of context with specific cultures seems to imply that the degree of context-dependence is merely the result of historical accidents or of idiosyncratic differences between ethnicities. To develop a more systematic, scientific understanding of the relative importance of context in different situations, we need to be able to measure context-dependence in a reliable, objective and accurate way. If we could make a quantitative estimate of the degree of context-dependence in a particular situation, then we would be able to determine how this degree covaries with different features of the situation. For example, using such a measure, we would be able to either prove or refute the above assumption that twins maximally rely on context when communicating with each other. Moreover, we could use such a measure to either suggest or test hypotheses about the fundamental factors that determine the amount of context-dependence.

The present paper will first examine in more depth the role of context in linguistic communication. This will allow us to define a fundamental dimension of variation between different linguistic styles, going from the high-context pole, which we will call "contextual", to the low-context one, which we will call "formal". By analysing the degree of context-dependence of words belonging to different grammaticalcategories, we will then develop an overall measure for the degree of contextuality or formality of a language excerpt. Using a variety of data from different languages, we will show that this measure accurately distinguishes the more contextual from the more formal genres. Finally, we will examine a number of external factors that affect the degree of contextuality of a communicative situation, confronting theoretical hypotheses with the empirical data that are as yet available. Thus, our approach fits in with the "grassroots" approach to context, which starts from the observation of concrete phenomena rather than from *a priori* abstractions. This will allow us to put context

into context, that is, examine the features of the wider situation that determine how important context is in any given communication.

2. Formality versus contextuality in language

In order to minimize ambiguity and maximize the objectivity and universality of its statements, science tries to express its result as much as possible through formal languages (Heylighen, 1999). This is necessary in particular for models that are to be implemented as computer programs. Artificial Intelligence can be defined as an approach that tries to develop computational models of human cognition and communication. To achieve this, AI makes use of various formal languages, such as predicate logic or semantic networks. However, recent developments have made it clear that complete formal representation is not only theoretically, but also practically impossible, and that AI systems will have to take into account the context in which they use their models (AAAI, 1997). A major source of inspiration for this shift from closed, formal models to context-dependent ones is natural language, where context enters the interpretive process from the very beginning.

It is a commonplace that natural languages, such as English, are very different from formalisms. However, Grice's (1975) classic paper on "Logic and Conversation" sets out to show that the divide is not as deep as one tends to believe. Much of what in a formal language must be expressed explicitly in order to avoid ambiguity, will be conveyed in natural language by *implicature*, that is, by implicit reference to a shared framework of knowledge and its implications. For example, if a person entering a room with an open window through which wind is blowing says "It is cold here", the likely implicature is "I would like the window to be closed". Though that message was not uttered literally, it is easily inferred from the background knowledge that heated rooms become warmer when windows are closed, and that people prefer not to feel cold. Grice (1975) points out that if one takes into account this shared context (including the general rules or "maxims" of conversation), expressions which appear ambiguous or non-sensical when interpreted on their own become clear and logical.

The conclusion is that natural language will appear much less ambiguous and more logical than it might have seemed if one takes into account different unstated background assumptions. What really sets formal languages apart is the fact that they try to achieve the same clarity *without* unstated assumptions. In order to analyse this further we must examine the essential role of context in resolving semantic ambiguity (cf. Gorfein, 1989) and in understanding linguistic structure (cf. Duranti & Goodwin, 1992).

This role can be illustrated most clearly by considering simple expressions, that must be anchored, or attached, to some part of the spatio-temporal context in order to be meaningful. Such anchoring is called *deixis* (see e.g. Levelt, 1989: 58). Examples are simple expressions like "I", "his", or "them", which must be connected to a particular person, "here", "over there", or "upstairs" which must be attached to a particular place, and "before", "now", or "tomorrow", which must be linked to a particular time. Deictic words on their own have a variable meaning. "He" might refer to John Smith, to Peter Jones, or to any other male member of humanity. Yet, only one of them will

be referred to in any actual expression. Which person that is will be determined by the context.

We will use the general term context-dependent or *contextual* for expressions such as these (cf. Dewaele, 1995), which are ambiguous when considered on their own, but where the ambiguity can be resolved by taking into account additional information from the context (cf. Heylighen, 1999). In philosophy, such expressions are usually called "indexical" (Bar-Hillel, 1954; Barnes & Law, 1976). The term "contextuality" encompasses both the case of deixis, where a connection is to be made with a concrete part of the spatio-temporal setting, and the more abstract case of implicature, where the information to be added must be inferred from unstated background assumptions. It also includes reference to information expressed earlier, which is called "anaphora" in linguistics. More generally, the *context* of an expresssion can be defined as *everything available for awareness which is not part of the expression itself, but which is needed to correctly interpret the expression*.

The opposite of contextuality may be called "formality". Formal language will avoid ambiguity by including the information about the context that would disambiguate the expression into the expression itself, that is to say, by explicitly stating the necessary references, assumptions, and background knowledge which would have remained tacit in a contextual expression of the same meaning.

For example, the contextual expression "I'll see him tomorrow" can be rephrased more formally as "Karen Jones will meet John Smith on October 13, 2001". For somebody who knows the context, i.e. who knows that the speaker is Karen Jones, that she is thinking about John Smith, and that today is October 12, 2001, the two sentences contain exactly the same amount of information. But someone who does not know the context—for example a person who read the sentence on a piece of paper, not knowing who wrote it or when that happened—would find the second sentence much more informative.

The choice between the two ways of formulating the same idea will clearly depend on how much knowledge the persons to whom the message is addressed are presumed to have about the context in which it was uttered. The less they know, the more important it is to avoid contextual expressions, replacing them by explicit characterizations. On the other hand, when the audience has a good knowledge of the context, there is a clear advantage in using contextual expressions, such as "I", "him" or "tomorrow", which are shorter and more direct. This can be illustrated by considering the following sequence of increasingly formal descriptions of the same person: "he", "John", "John Smith", "Dr. John K. Smith, assistant anaesthesist at the neurology unit of St. Swithin's hospital". Each term in this sequence is less dependent on the context for its correct interpretation, but correspondingly longer, than the previous one. Which level of formal specification is chosen will depend on Grice's (1975) maxim of quantity: the message should be as informative as is required, but not more.

Let us summarize the main reasons why someone would prefer formal expressions to contextual ones, or vice-versa (Dewaele, 1995; Heylighen, 1999). The basic advantage of formality, which follows from its definition, is that more formal messages have *less chance to be misinterpreted* by others who do not share the same context as the sender. This is clearly exemplified by written language, where there is no direct contact between sender and receiver, and hence a much smaller sharing of

context than in speech. We have demonstrated that written language is in general more formal or explicit than spoken language (Dewaele, to appear a). The definition also implies that validity or comprehensibility of formal messages will extend over wider contexts: more people, longer time spans, more diverse circumstances, etc. This makes it easier for formally expressed knowledge to maintain and spread over many different persons, groups or cultures (Heylighen, 1993, 1999).

The concurrent disadvantage of invariance over contexts is that formal speech is more static or *rigid*, and will less easily accommodate to phenomena that demand expressions with a meaning different from the one found in dictionaries. Contextual speech, by definition, is *flexible*: meanings shift when the context changes. This is particularly useful when phenomena are to be described for which no clear expression is available in the language as yet. By using eminently contextual expressions like "it" or "that thing there", it is possible to refer to the most unusual phenomena.

The second disadvantage of formal speech is that it is structurally more complex. Therefore, formal expressions require more time, attention and cognitive processing to be produced and understood. The absence of context, as Givón (1985) observed, forces the language user to code the necessary presuppositions within the message. The resulting "syntactic mode" (Givón, 1985: 1018) of expression involves a higher use of nouns that require more lexical searching because of their relatively infrequent use. Contextual speech, on the other hand, can do the job with less, shorter, and more frequent words, which are easily and quickly retrieved, and less need for precision, since the context shared by sender and receiver will provide the additional information lacking in the linguistic expression itself. Non-verbal communication can, moreover, help dissolve ambiguity. (Givón (1985) calls this contextually rooted language "the pragmatic mode".)

Contextual speech-styles will also be more *interactive* or *involved*, reacting immediately to the interlocutors, events or other elements of the context, rather than describing things from a detached, impersonal, "objective" point of view.

The conclusion is that the degree of contextuality of an expression will depend on the requirements of the situation, but that there will still be an element of personal choice, depending on whether the sender prefers accuracy over flexibility, detachment over involvement, or fears possible misinterpretation more than additional cognitive load. As a general rule we will expect contextuality to be lowest in the more static, intellectual or informational forms or expression, such as legal or scientific documents, and highest in the more interactive and personal communication situations, such as conversations or personal letters.

The most reliable way of studying these dependencies is by empirical observation, where expressions produced in different situations or by different subjects are compared as to their overall contextuality, in the hope of finding recurrent relationships. In order to research such dependencies, however, we must first devise an empirical measure for contextuality.

3. Measuring language contextuality

3.1. Word category frequencies and the F-measure

Although the above theoretical definition of contextuality appears intuitively adequate, one might wonder whether it is possible to extend it to some practically useful and reliable measure that would allow an observer to distinguish more contextual from less contextual discourses. Such a measure should be both *valid*, in the sense that what it measures effectively corresponds to contextuality as it was defined and as it is intuitively understood, and *practical*, in the sense that it does not require an inordinate amount of effort to apply. These two criteria are inherently at odds: the more valid a measurement needs to be, the more precise and detailed the procedure will be, and the more time and effort will be invested in carrying it out.

The measure we wish to devise should offer a good compromise between these two requirements. Its procedures should be easy to apply to large corpora of linguistic data, without requiring specific rules for handling all possible subtleties or exceptions of the particular language or situation. Yet, it should be capable to unambiguously distinguish discourses that are considered formal from those that are considered contextual.

Determining an average degree of contextuality seems more easy when focusing on cases of deixis or anaphora at the level of single words rather than contemplating complex implicatures at the level of sentences and situations[1]. Analysing language at the level of the lexicon makes it possible to avoid all intricacies at the level of phonetics, syntax, semantics and pragmatics. The analysis of the numbers and types of words in a text is quite easy to automatize by means of computer programs. In contrast, recognition of phonetic patterns, syntactical parsing, and even more semantic and pragmatic interpretation of natural language are still very difficult—if not plain impossible—to perform automatically.

Our basic approach is to divide the words of the lexicon into two classes, depending on whether they are used mainly to build more context-dependent or more context-independent speech. In the one class, we will list all words with a deictic function, i.e. that require reference to the spatio-temporal or communicative context to be understood. Levelt (1989: 45) distinguishes four types of deixis: referring to person ("we", "him", "my",...), place ("here", "those", "upstairs",...), time ("now", "later", "yesterday", ...), and discourse ("therefore", "yes", "however", ...). The latter category of deixis includes anaphora: reference to things expressed earlier. Further examples of discourse deixis are exclamations or interjections like "Ooh!", "Well", or "OK". In

---

[1] A preliminary investigation by Mazzie (1987), extending work by Prince (1981), concluded that the relative proportion of “evoked” contextual information (deictic or anaphoric, directly referring to contextual elements) versus “inferrable” contextual information (indirectly derived, e.g. by implicature) did not depend on the mode of expression (written vs. spoken) but only on its content (abstract vs. narrative). It would be interesting to check in how far this result can be generalized to corroborate our simplifying assumption that evoked contextuality is a good measure of overall contextuality, and thus of formality.

logic, deictic and anaphoric words would correspond to *variables*, which do not have a fixed referent or interpretation[2].

In the other, non-deictic, class are the words referring to an intrinsic class of phenomena, which does not normally vary under changes of context. These would correspond in logic basically to *predicates*. Examples are most nouns and adjectives (e.g. "tree", "women", "red", ...).

Ideally, a measure of formality would start from a classification in which an average degree of deixis would be attributed to every word of a language (cf. Leckie-Tarrie, 1995). The contextuality of a text could then be determined by calculating the total deixis averaged over all of its words. The development of such a classification, however, would be a very long and intricate task, which would have to be started from scratch for every new language.

A much simpler, but coarser, measure can be developed by determining an average degree of deixis not for individual words but for the conventional grammatical categories of words. Our examples of contextual words belong basically to the categories of pronouns, adverbs and interjections. Pronouns are particularly clear examples of deictic words. Typically context-independent words are nouns, adjectives (which further specify the meaning of nouns) and prepositions (which mainly create a relation introducing a noun phrase with additional information).

Although verbs seem to function as predicates, and might therefore seem similar to the non-deictic nouns, inflected verbs are intrinsically deictic because they refer implicitly to a particular time through their tense (time deixis, cf. Levelt, 1989: 55), and to a particular subject through their inflection (person or object deixis). The latter feature is especially important in languages like Spanish, Latin and Italian, where a pronoun does not have to be stated as a subject of the sentence, since it can be inferred directly from the inflection of the verb. This makes an expression using an inflected verb much more contextual than an equivalent expression without the verb.

This can be illustrated by eliminating deixis from a simple sentence like "They destroyed a building". Removing person deixis, we get the more formal, passive expression: "A building was destroyed". In order to further remove time deixis, we must replace the verb by a noun (this is called "nominalization"): "The destruction of a building". The latter phrase is much less contextual, but correspondingly more static, detached and impersonal. It might be used to express an abstract or general rule (e.g. "The destruction of a building is a dangerous activity") rather than a specific event taking place in a given context, like the original phrase.

Apart from simple exclamations ("You there!"), it is impossible to build sentences without verbs or nouns. Since verbs and nouns are to a certain degree interchangeable (by nominalization or its inverse, verbalization), it will depend on the speaker whether he or she will primarily use verbs or nouns as means of expression. Given the fact that (inflected) verbs are necessarily deictic, whereas nouns are not, we may assume that a speaker using a formal style will prefer nouns (cf. Halliday, 1985), while a speaker using a contextual style will prefer verbs. This increase in verb proportion in

---

[2] In fact there exists at least one programming language (HyperTalk) in which certain variables are used in a way similar to deictic words in natural language: e.g. “it” refers to the last expression put in memory, “me” refers to the object that is performing the command.

contextual styles will be reinforced by the fact that the more formal noun phrases, including nouns, articles, adjectives and prepositions, used to specify additional details about the context, will tend to be left out completely or replaced by pronouns without further determiners.

Verbalization/nominalization of phrases will normally also transform adjectives into adverbs, or vice versa. Thus, the frequency of adverbs will increase with an increase in verb frequency, and decrease with an increase in noun/adjective frequency. This puts adverbs indirectly (via their connection to verbs) in the deictic category, although they might otherwise seem similar to the predicative adjectives, both categories expressing attributes added to other words (nouns, adjectives or verbs). Moreover, the most frequent adverbs have a direct deictic function: e.g. "thus", "yes" (discourse deixis), "later" (time deixis), or "there" (place deixis). In that use, they are similar to possessive or demonstrative pronouns ("mine", "this", etc.).

Although articles ("a", "the") might seem related to demonstrative pronouns ("this", "that"), Kleiber (1991) argues convincingly that they are non-deictic. Moreover, their frequency for obvious reasons covaries with the one of nouns. Therefore, they may be put in the non-deictic class.

Conjunctions, which have no reference, neither to an implicit context, nor to an explicit, objective meaning, do not seem to be related to the deixis or formality of an expression, but only to its structure. Therefore, they are not put into either category (cf. Dewaele, 1996a, 1996b).

In conclusion, the formal, non-deictic category of words, whose frequency is expected to increase with the formality of a text, includes the *nouns*, *adjectives*, *prepositions* and *articles*. The deictic category, whose frequency is expected to increase with the contextuality of a discourse, consists of the *pronouns*, *verbs*, *adverbs*, and *interjections*. The remaining category of conjunctions has no a priori correlation with contextuality. If we add up the frequencies of the formal categories, subtract the frequencies of the deictic categories and normalize to 100, we get a measure which will always increase with an increase of formality. This leads us to the following simple formula:

F = (noun frequency + adjective freq. + preposition freq. + article freq. – pronoun freq. – verb freq. – adverb freq. – interjection freq. + 100)/2

The frequencies are here expressed as percentages of the number of words belonging to a particular category with respect to the total number of words in the excerpt. F will then vary between 0 and 100% (but obviously never reach these limits). The more formal the language excerpt, the higher the value of F is expected to be.

Although the subcategories (nouns, verbs, etc.) are here listed explicitly, the formula can be made more general by just adding whichever words seem the more formal and subtracting whichever words seem the more deictic. This is useful in situations where the above grammatical categorizations are ambiguous or where data are lacking (e.g. the number of nouns might be known, but not the number of articles or interjections). As long as there are sufficient words in each of the two supercategories, the resulting measure should be sufficient to distinguish different

degrees of contextuality. The practical effectiveness of this measure will now be illustrated by applying it to data from different languages.

3.2. Application of the F-measure to data

A number of studies by one of us (Dewaele, 1995, 1996a, 1996b, in press a,b), on the use of advanced French interlanguage in different situations, provides extensive data about frequencies of different word categories. A corpus of 2 speech-styles and 1 written style was collected from a group of students in three situations, in decreasing order of contextuality: 1) an informal conversation; 2) an oral examination, testing the subject's knowledge of the language; 3) an essay produced during a written examination. In agreement with our above predictions, the frequency of nouns, adjectives, articles and prepositions increased with an increase of formality in the situation, while the frequency of pronouns, adverbs and verbs decreased. The frequency of conjunctions had no special relation with contextuality. This led to values for the F-scores of respectively 44 (informal), 54 (examination) and 56 (essay)[3].

These results could be interpreted as a mere peculiarity of interlanguage or of exam situations. More general data about word frequencies for different languages and situations are available, however. After an analysis of frequency dictionaries of Italian and Dutch, some data about word categories in English, and a small corpus of French, we found similar variations of word frequencies between more and less contextual styles. Written language scores much higher on the F-measure than spoken language (Dewaele, in press a), as could be expected from the fact that one can rely much less on shared context in writing than in speaking.

For the Dutch list of frequencies of Uit den Boogaert (1975), which seemed the most reliable (frequencies based on a total of about 120 000 words per genre), we get an average F(written) = 62, F(spoken) = 42. More specifically, word frequencies taken from more informational genres, such as scientific texts (F=66) or (broadsheet) newspapers (F=68), lead to much higher formality scores than those from more involved genres like novels (F=52) or family magazines (F=58) (Uit den Boogaert, 1975). Within spoken language, the speech of people with an academic degree (F=44) not surprisingly scores higher than the one of people without an academic degree (F=40) (calculated on the basis of data from Uit den Boogaert, 1975), and, less obviously, that of men (F=42) higher than that of women (F=39) (calculated on the basis of data from De Jong, 1979). The general ordering agrees quite well with intuition as to which genres are the more formal. The formality scores for different sources in Dutch are summarized in Table 1.

	formal categories	*contextual categories*	

[3]The relatively small difference in formality between the written and spoken formal situations might be explained by the specificity of the interlanguage situation: the limited vocabulary in the second language will tend to restrict the higher precision of expression which would otherwise be expected for written essays.

	Nouns	Articles	Prepos.	Adject.	Pronouns	Verbs	Adverbs	Conj.	**Forma-lity**
Oral Female	10.4	6.9	5.9	8.1	17.0	19.4	17.5	7.5	**38.7**
Oral N.Acad.	12.8	8.5	6.3	6.7	16.0	18.8	19.3	6.3	**40.1**
Oral Male	11.5	8.2	6.7	7.6	15.8	18.5	16.5	7.1	**41.6**
Oral Acad.	13.2	9.6	7.9	7.1	14.0	17.8	17.9	7.1	**44.1**
Novels	18.5	10.5	10.3	10.0	13.3	20.6	10.5	6.1	**52.5**
Fam. Magaz.	21.8	9.8	12.2	11.1	10.1	18.7	9.7	6.4	**58.2**
Magazines	24.2	11.6	13.9	10.9	8.6	17.7	8.7	4.3	**62.8**
Scientific	23.1	15.0	13.8	10.8	6.7	16.6	8.0	6.0	**65.7**
Newspapers	26.0	14.7	14.5	10.6	5.6	16.7	7.2	4.7	**68.1**

Table 1: frequencies in percents and resulting formality scores for Dutch language coming from different fields (words for which the category is unclear or ambiguous were left out, so that the frequencies do not add up to 100%.)

When we look in more detail at the frequencies of the separate word categories (Table 1), we notice that the frequency of the "formal" categories (nouns, articles, adjectives, prepositions) increases with an increase of formality, while the frequency of the "contextual" categories (pronouns, verbs, adverbs—data on interjections are not available for all genres) decreases, except for one or two outliers per category. This confirms our hypothesis that these categories increase or decrease together when the style becomes more formal, but that the overall effect captured in the F-score is more reliable than any single category. The frequency of the conjunctions, on the other hand, does not clearly increase or decrease. (the tendency towards decrease in the Dutch sample is counterbalanced by a slight tendency towards increase in our advanced French interlanguage data, and an almost constant trend for the Italian data).

When comparing the individual categories, we note that the pronouns (decreasing) are the only ones moving monotonically with formality. This could be expected since pronouns form the most clearly contextual category, which might therefore be expected to correlate best with formality. Verbs, on the other hand, decrease rather slowly and irregularly, perhaps signalling their dual predicative/non-finite and deictic/finite nature. Within the "formal" categories, prepositions perform best. This becomes less surprising if we note that prepositions are typically used to start a further specification, replacing a direct reference to the context (e.g. replacing "there" with "*on* the table", or "afterwards" with "*after* the dinner"), or simply adding precise information on the circumstances in which something happens.

On the basis of the frequency dictionaries of Bortolini et al. (1971) [A], and of Juilland & Traversa (1973) [B], we made similar calculations for Italian. The ordering of genres we get is remarkably similar to the one for Dutch, except for a reversal of the positions of the "scientific" and "newspaper" sources, which may be due to a different way of selecting the sources. Language used in Italian movies and theatre (which is

supposed to approximate every-day speech) has formalities of 48 (A) and 52 (A) or 53 (B) respectively. Novels, depending on the sample chosen, score 58 (A) or 64 (B). Newspapers and magazines score 66 (A) or 71 (B). Essays, and Technical and Scientific Writings, (both B) score respectively 69 and 72 (see Table 2).

We notice a clear difference between the two dictionaries, the samples from B scoring systematically higher than the corresponding samples from A. This is probably due to the way the data were collected, including definition of the word categories and selection of the samples. A systematic difference is that the corpora used for B date from before the 2nd World War, while the ones used for A date from after the war. This might signify that a less formal writing style developed in more recent periods.

	formal categories				*contextual categories*					
	Nouns	Art.	Prep.	Adj.	Pron.	Verbs	Adve.	Interj.	Conj.	**Formality**
Movies A	13.4	8.3	8.6	5.1	1.6	27.0	10.0	0.8	6.0	**48.0**
Theatre A	14.8	10.2	9.4	5.5	1.4	24.5	8.7	0.8	5.6	**52.3**
Theatre B	14.0	10.2	10.5	4.8	1.4	23.9	8.1	0.1	7.2	**53.0**
Novels A	16.7	13.8	14.0	5.6	8.5	20.1	6.5	0.1	6.4	**57.5**
Novels& Sh.Stories B	18.2	16.0	15.5	6.7	7.0	17.7	4.5	0.1	6.3	**63.6**
Newspapers A	18.9	16.8	16.7	7.7	5.1	17.5	4.9	0.0	5.2	**66.3**
Essays B	19.0	16.9	17.2	8.1	5.8	12.9	4.2	0.0	7.0	**69.1**
Newspapers& Magazines B	20.4	18.4	18.4	8.4	4.3	15.4	3.5	0.0	5.3	**71.2**
Technical& Scientif. B	18.6	18.0	20.2	7.6	4.3	12.7	4.1	0.0	6.0	**71.6**

Table 2: frequencies in percents and resulting formality scores for Italian language coming from different fields (words for which the category is unclear or ambiguous were left out, so that the frequencies do not add up to 100%.)

When we look at word categories, we again see results very similar to the ones for Dutch, except for one complicating factor: subject pronouns in Italian do not have to be stated explicitly, as the referent can be inferred from the form of the verb. As a result, the frequency of pronouns does not correlate well with the other formality components, since the absence of a pronoun does not imply the presence of a noun. Still, the other components, and in particular the verbs, seem to make up for this effect by even stronger correlations with formality. This may be due to the fact that the removal of pronouns as subjects of the phrase puts the burden of person deixis wholly on the verb. The relatively small number of pronouns may also explain the higher overall formality scores of Italian when compared to Dutch. The categories best correlating with F seem to be the prepositions (confirming their role in Dutch) and the

interjections (which were not used in our calculations for Dutch). The overall frequency of interjections is very small, though, so that their effect is not very important.

It is interesting to note that Zampolli (1977) performed different statistical analyses (Chi^2, Z, ...) on these same data about word categories from the two Italian frequency dictionaries. He found the same unequivocal mathematical ordering of the different genres, and calculated that the probability of this ordering being due to chance is virtually zero. However, he concluded by regretting the lack of any theory that could offer an adequate explanation of these results. It seems that our present concept of formality/contextuality would answer Zampolli's questions.

Hudson (1994), in a similar reflection about the proportions of word classes in the data he gathered (mostly for English), comes to the following conclusion:

> there seem to be regularities in language of which most of us have been completely unaware - regularities which involve the statistical probability of any randomly selected word belonging to a particular word-class. At present we have no hope of explaining these regularities, but they are a challenge that our grandchildren may (possibly) be able to meet (Hudson, 1994: 337).

Again, a large part of his questions can be answered by our theory of contextuality. Although Hudson's data are less detailed than the data used by Zampolli (lacking frequencies for several of the word classes), the data from his table 6 for written and spoken English are sufficiently elaborate to apply a simplified formality measure, F* (where the star denotes the absence of numbers for the article and interjection categories). The results are shown in Table 3 .

	formal categories			contextual categories			
	Nouns	Prepos.	Adject.	Pronouns	Verbs	Adverbs	**Forma-lity***
Phone conversations	14	7	4	17	25	11	**36**
Conversations	15	8	4	16	24	11	**38**
Spontaneous speeches	18	9	5	15	21	9	**44**
Interviews	18	11	6	13	21	10	**46**
Imaginative writing	22	10	6	15	22	7	**47**
Prepared speeches	21	11	5	11	19	8	**50**
Broadcasts	24	12	6	7	14	12	**55**
Writing	28	12	7	9	18	5	**58**
Informational writing	30	13	8	7	17	5	**61**

Table 3: formality (lacking frequencies of some word categories) scores for English language coming from different fields.*

Again, we note that the formal categories mostly increase together with formality, while the contextual categories decrease, and that the ordering of genres according to

formality corresponds quite well with intuition and with expectations based on our theoretical model (although it is not clear why the phone conversations would be more contextual than the face-to-face conversations). From Hudson's other data, the only ones elaborate enough to allow a comparison of formality measures are the data from New Testament Greek, where the higher formality of the letters compared to the narrative follows the same pattern as the one between informational and imaginative genres in written English, and the data from children's English, where the free play excerpts are markedly more contextual than the interviews, and where boys' language is less contextual than girls' language.

Finally, as an additional check, we analysed a few samples of French. A television interview with a call-girl scored 45, an interview with the president of the republic scored 52, an address to the nation by the president scored 58, and an article in an intellectual newspaper scored 78, confirming the general tendencies observed for English, Dutch and Italian.

3.3. Formality as a universal factor

In spite of these empirical confirmations, our definition of F may seem to some degree arbitrary, just another one of many related, but different, dimensions proposed by different authors, which all correlate to some degree with variations such as written vs. spoken, but whose underlying motivation is debatable. We will now attempt to show that a dimension akin to formality appears like an inevitable outcome of any in-depth analysis of linguistic variation.

In the previously mentioned studies on French interlanguage (Dewaele, 1995, 1996a, 2000, in press a, b) a variable similar to the F measure automatically emerged from a principal components factor analysis conducted on the proportions of word categories between different samples of language, produced by different subjects in a similar situation. All samples were characterized by their values on 7 variables, representing the frequencies of the following word categories: nouns, determiners (articles + adjectives), prepositions, verbs, pronouns, adverbs, and conjunctions. Factor analysis is a statistical technique which attempts to reduce the variation between the samples to a minimal number of newly derived components or factors. The resulting factors are linear combinations of the original variables. First the combined variable is selected that explains the highest amount of variance, then the one with the second highest variance, and so on, until the remaining variation becomes too small to be significant.

For each of two situations (informal conversation, formal oral examination), a separate factor analysis was performed. Each time, two main orthogonal factors appeared. The first one, which explained over 50% of the variation, was called explicitness/implicitness. It is practically identical to formality/contextuality as we have defined it, since nouns, determiners and prepositions obtained strong positive loadings on this factor, whereas pronouns, adverbs, and verbs obtained strong negative loadings. The second factor, explaining between 10 and 20% of the variation, shows only weak correlations with the different frequencies, except for the one of the conjunctions. It was therefore interpreted as a measure of the complexity of sentence structures, independent of their degree of formality (cf. Dewaele, 1995).

In conclusion, even if we do not compare situations or genres with different external requirements of formality, there appears a stylistic variation between samples that very closely mirrors our definition of the contextuality variable. This variation is apparently due to the personal preferences of the subjects for more or less contextual styles of expression. Moreover, this variation—at least at the level of word categories—is by far the most important one, explaining more than half of the variance between samples.

This result is further strengthened when a similar factor analysis is performed with the above-mentioned data (tables 1 and 2) of word frequencies for different genres (unfortunately, the number of genres is too small for a reliable factor analysis), in each of three languages, Dutch, Italian and French. The results are quite similar, except that the variance explained by the first factor, formality, is even greater: from 70% (for French, where the samples were very limited) to over 80% (for Italian and Dutch). A likely cause is that the samples were more diverse in situational formality than the samples in the former study, which were all produced in similar (formal or contextual) situations.

A very extensive factor analysis of different styles in English by Biber (1988) confirms these general results. He starts with a long list of linguistic variables, including fine-grained word categories (e.g. private verbs, 2nd person pronouns, place adverbials), but also different grammatical and stylistical features, some of which are typical for English (e.g. do as proverb, number of agentless passive sentences, contractions, that clauses as relative complements, etc.). His analysis produces 7 factors. The first one, an extremely powerful factor representing a very basic dimension of variation among spoken and written texts in English (Biber, 1988: 104) is very similar to our definition of contextuality. This factor, which Biber calls "involved versus informational production", correlates positively with the most frequent verb and pronoun forms, with adverbs and different types of interjections. It correlates negatively with nouns, prepositions and attributive adjectives.

Biber's interpretation of the factor seems compatible with our analysis, except that he has some difficulty fitting the empirically derived factor into a single theoretical construct. He rather distinguishes two separate parameters (Biber, 1988: 107): on the one hand, precision and density of information; on the other hand, interaction, involvement and affection. He proposes a not very convincing explanation why these *a priori* independent dimensions are negatively correlated, by noting that involved situations, such as conversations, tend to be characterized by time pressure, which makes it difficult to achieve high precision. This forces him to paradoxically explain the low precision characterizing personal letters by self-imposed time constraints (Biber, 1988: 108). In our analysis, both involvement and lack of precision are characteristic of a contextual style of expression, where references to the shared context both signal close contact or involvement, and obviate the need for a precise description of that context. In this view, personal letters lack detailed expositions not because of time pressure (composing letters can take as much time as desired), but because the intimately known person to whom the letter is addressed is assumed to already know the details about the context in which one is writing.

The scores of different genres of language on Biber's factor 1 also confirm our results (cf. Table 3, based on Hudson's (1994) reprocessing of part of Biber's original

data). Ordered from the most involved genres to the most informational ones, we get: telephone and face-to-face conversations; personal letters, spontaneous speeches and interviews; different types of fiction, prepared speeches, professional letters and broadcasts; biographies, academic prose and press reportage; and finally official documents, which score lowest of all on involvedness (see also Biber, Conrad and Reppen, 1994: 182). This ordering seems to reflect expectations based on either intuition or our theoretical analysis of contextuality. Our application of the F-measure to (part of) the same data (Table 3) produces an identical ordering of genres, however, with a much smaller effort of analysis, a clearer interpretation, and an easier generalization to other languages.

In later work, Biber extends his factor analytic methodology to the very different languageof Somali (Biber and Hared, 1992), and compares the results with similar studies of Korean (Kim and Biber, 1995) and Nukulaelae Tuvaluan (Besnier, 1988), a language spoken by a few hundred people on a Polynesian atoll. In all three cases, the same involved versus informational factor as in English comes out markedly as the strongest dimension of variation between registers. It is variously called "involvement versus exposition" (Biber and Hared, 1992), "interaction versus information" (Besnier, 1988), and "informal interaction versus explicit elaboration" (Kim & Biber, 1995). Adding our results on Dutch, French and Italian, this brings us to a total of seven languages, belonging to four completely different language families, which all appear to share the same fundamental dimension of variation, captured by our concept of contextuality/formality.

Of course, as Biber notes (1988), no single variable can represent all types of variation between genres or registers. Between 3 and 7 major dimensions came out of the four factor analytic studies reviewed by Biber and Hared (1992). However, only the involved-informational factor was shared by all samples, while the less strong narrativity factor (characterized by the use of past tense and third person) was shared by all samples except the Tuvaluan (possibly because of insufficient data). The remaining factors seemed to reflect specificities of the different languages. It is hard to avoid the conclusion that a dimension similar to contextuality appears as *the* most important and universal feature distinguishing styles, registers or genres in different languages.

4. Non-linguistic determinants of contextuality

As the contextuality concept appears both theoretically and empirically to be well-defined, the time seems ripe to test its predictive and explanatory power in practical situations. We will now examine some non-linguistic variables that affect the degree of contextuality. This degree will in the first place be determined by the characteristics of the situation in which the linguistic behavior was produced, and by the psychological characteristics of the speaker. Both situation and personality are complex, multidimensional phenomena. In the following we have limited the list of factors that may affect contextuality to those variables for which we have some empirical evidence, and a (preliminary) theoretical interpretation.

4.1. Situation

We defined formality as avoidance of ambiguity in order to minimize the chance of misinterpretation. This means, first of all, that formality will be highest in those situations where accurate understanding is essential, such as contracts, laws, or international treaties. This may explain the very high formality of official documents according to the data from Biber (1988). It also explains why in our French interlanguage experiment, the oral exam scored much higher on formality than the relaxed conversation.

Second, formality will be higher when correct interpretation is more difficult to achieve. One way to secure accurate understanding is corrective feedback: if the listener can signal to the speaker when he or she doesn't understand, so that the speaker can reformulate the phrase, the speaker will need to worry less about unambiguous expression. Thus, conversations require less formality than speeches or than written texts (cf. table 3). Within written language, letters, which normally expect a reply, will be more contextual than articles or books, without possibility for reply, as confirmed by the data from Biber (1988). This also fits in with Gudykunst & Ting-Toomey's observation (1988) that in a low-context culture the burden of communication is placed on the sender, whereas in a high-context situation, communication is much more interactive, involving both sender and receiver.

The most important determinant of the probability of misinterpretation, though, is the context shared by sender and receiver of a message. We could summarize an act of communication or transfer of information by the following formula: E + C → I, where E stands for the expression produced by the sender, C for the context shared by sender and receiver, I for the interpretation by the receiver, and the arrow for determines (cf. Heylighen, 1999). The larger C, the smaller E can be, and therefore the lower E's formality. The smaller the size of the shared context, though, the more information needs to be put into the expression in order to make sure that all information intended by the sender effectively reaches the receiver.

The number of elements in the context is potentially infinite: any characteristic of the physical, social and mental situation can influence the interpretation of an expression. However, in order to simplify the analysis, we will limit ourselves to the most basic dimensions. Following Levelt's (1989) classification of linguistic deixis, we can distinguish four categories of context factors: the *persons* involved, the *space* or setting of the communication, the *time,* and the *discourse* preceding the present expression. The general principle that a decrease in shared context leads to an increase in formality can now be used to produce specific predictions for each of these dimensions.

The persons involved are in the first place the sender and the receiver of the message. All other things being equal, the larger the difference in psychological or cultural background (including characteristics such as age, class, nationality, or education) between these interlocutors, the smaller the shared context, and therefore the higher the formality of their communication. This may explain the requirement of politeness, characterized by a formal style of language that uses more nouns (Brown & Levinson, 1979), when addressing strangers or people of a different rank. On the other hand, people who are psychologically close, such as siblings, spouses or intimate friends, will tend to be minimally formal in their exchanges. Following Hall

(1976), we would hypothesize that the highest degree of contextuality will be found among identical twins that were raised together, who completely share their cultural, social and even biological backgrounds. More generally, we can assume, together with Hall (1976), that high contextuality will be found primarily in environments where there are strong social ties between the participants, and where there is a high level of mutual knowledge, shared experience and commitment. This explains why the USA, where people travel a lot, have many, short-term relationships, and diverse cultural backgrounds, is a typical example of a low-context environment. Japan, on the other hand, where culture is much more homogeneous and social bonds are much stronger and more rigid, is a typical high-context environment (Hall, 1976). Given the present trend of globalization, where the number of contacts with people from different backgrounds increases, whereas the duration of relationships tends to decrease, we might predict that the average contextuality of communication will tend to decrease all over the world. Another implication of this model is that contextuality shifts within a personal relationship will signal changes in intimacy: an increase in contextuality indicates a warming of the relationship, whereas a (more unlikely) decrease communicates distancing or unease, implying that expectations have not been met.

The study of Fielding & Fraser (1978) on interpersonal interaction indeed found that speech addressed to a liked listener is significantly less nominal (formal) than speech addressed to a disliked person. A further confirmation comes from Biber's (1988) analysis, which finds personal letters (addressed to a well-known person) to be markedly more involved (contextual) than professional letters. Our study of French interlanguage (Dewaele, 1993a, 1996a, 1996b, in press a,b) provides some further evidence. The subjects (university students) were classified on a four point scale measuring social background, depending on whether their parents finished their education after junior secondary school, senior secondary school, non-university higher institute, or university. The formality of their language correlated negatively with the parents' educational level. This might be explained by assuming that the interviewer (a university assistant) was viewed as more distant on the sociocultural level by the subjects whose parents came from a lower educational background.

Another implication of our model concerns audience size. All other things being equal, the larger the audience, the less the different receivers and the sender will have in common, and thus the smaller the shared context. Moreover, the larger the audience, in general, the more important it will be to secure accurate understanding. Therefore, we may expect that speeches or texts directed to a large audience will be more formal than comments addressed to one or a few persons. This is confirmed by the higher formality score of speeches compared to conversations, of broadcasts compared to speeches (see table 3), and of published texts compared to letters (Biber, 1988). A more detailed method to test this hypothesis would consist in gathering texts of speeches delivered to different audiences, and trying to correlate the formality score of the language with the size of the audience.

The more different the *spatial setting* for sender and receiver, the smaller the shared context. Therefore, conversations over the telephone or another indirect medium would be expected to be more formal than conversations which take place in the same location. Fielding & Cooper (1976) found that conversations over the intercom are more nominal (formal) than face-to-face conversations. Moscovici & Plon

(1966) found that speech becomes more nominal over the telephone or when conversants are put back-to-back, so that they cannot see each other. Biber's (1988) data (table 3) do not confirm this result: telephone conversations get a slightly more contextual score than face-to-face conversations, but this may be due to the fact that Biber's telephone data came from a quite different source than his conversation data.

The longer the *time span* between sending and receiving, the less will remain of the original context in which the expression was produced. For example, reports written for archiving purposes will be more formal than notes taken to remember tomorrow's agenda. This may also in part explain why spontaneous speeches, produced on the spot, have a much higher contextuality than speeches prepared at an earlier moment (table 3). Another way to test this proposition empirically might consist in measuring the contextuality of messages sent through fast media (e.g. fax or electronic mail) versus slow media (e.g. postal mail). A message that can be expected to reach the addressee the same day should on average be more contextual than a message that takes several days to get through.

Finally, the factor of discourse deixis suggests that formality would be higher at the beginning of a conversation or text, because there is not any previous discourse to refer to as yet. Every document or conversation needs to set out its proper context before it can start using anaphoric expressions such as "therefore", "it", "him", etc. Although we have not analysed any data yet that could support this hypothesis, testing it seems straightforward: it suffices to collect a range of opening sentences or opening paragraphs from articles, speeches or conversations and compare their average formality with the formality of sentences from the middle of the same language sample.

Although we have discussed these different situational variables affecting formality separately, we must note that they are usually mixed in practical situations, which makes it more difficult to unambiguously test our hypotheses. For example, written and spoken language tend to differ in several of these aspects: sender and receiver of written texts are usually separated by time as well as by setting, and the possibility of feedback is usually much smaller than for speech. In this case, all differences point in the same direction, though: written language in general is less contextual than speech in general. This is confirmed by all the data we have reviewed. However, this does not mean that writing is always more formal than speech. For example, Biber (1988) found that broadcasted speech (e.g. radio or TV comments on live events, such as funerals or sports competitions), which is addressed to a very wide audience without possibility of feedback, is more "informational" (formal) than personal letters, addressed to one, intimately known person, who would be expected to respond. Some other situations we discussed depend on less variables. For example, the difference in formality between a presidential interview and a public address seems to reside mainly in the size of the audience, and the possibility for feedback.

4.2. Gender

There have been many studies of possible differences between the language of men and women, with interesting, but not easily interpreted, results. Though most

researchers find gender-related effects, there is some discussion on whether these differences are firmly substantiated (Thorne, Kramarae & Henley, 1983).

Our present data seem to indicate that women use a markedly more contextual speech style. On the basis of the Dutch frequency dictionary of De Jong (1979), we calculated a difference of 3 points on the F-measure between the sexes for speech (see table 1). These data are based on speech produced by 40 male and 40 female informants. A similar 3 point difference between male and female children's English is readily calculated from the data provided by Hudson (1995). The significance of these differences is confirmed by a more detailed statistical analysis of De Jong's data (Dewaele 2000), and by our study of advanced French interlanguage (Dewaele, 1996a, 1998, in press b).

In the latter study, the female part of the group scored F=39 on average in the informal situation, whereas the male group scored on average F=45, an overall difference of 6 points. In the formal examination situation and the written essays, no significant differences could be found, though. This seems to indicate that the influence of the situation is stronger than the effect of gender, which it overrides in those cases where spontaneous expression is more restricted. The difference in overall formality between formal and informal situations (10 points) is also clearly larger than the differences between genders within the same situation. The same pattern appears in the data from the Dutch frequency dictionaries (table 1), where the differences between genres are much larger than those between the sexes.

Let us try to interpret this apparent preference of women for more contextuality. From socio-linguistic and psychological studies (e.g. Hogg, 1985, Tannen, 1993, Coates 2000), it appears that women tend in general to be more intimate or *involved* in conversations, whereas men remain more distant or detached towards their conversation partners. Tannen (1993, 1992) concludes that men focus on the literal, informational content of the message, while women tend to focus on the implied relationship with their partner, an ill-understood difference in attitude, which creates many conflicts and misunderstandings between the sexes. As we argued earlier, involvement entails contextuality of the used language, since it implies direct and repeated reference to the people involved and to their previous reactions. This would lead, among other things, to more frequent use of pronouns, adverbs, inflected verbs and interjections. It also explains why the difference in contextuality between men and women was absent in the formal and written situations, where involvement is restricted for both sexes.

Tannen (1992) summarizes the stylistic differences between men and women by noting that the former are most comfortable with a style she calls report-talk, the latter with rapport-talk. Rapport-talk is aimed at building connection between the conversation partners and is most appropriate for what Tannen (1992) calls private speaking, involving conversations among couples or small, intimate groups. Report-talk functions to present objective information:

> Report-talk [...] does not arise only in the literally public situation of formal speeches delivered to a listening audience. The more people there are in a conversation, the less well you know them, and the more status differences among them, the more a conversation is like public

> speaking or report-talk. The fewer the people, the more intimately you know them, and the more equal their status, the more it is like private speaking or rapport-talk. (Tannen, 1992: 89)

Tannen's criteria for distinguishing the private and public situations are practically identical to the person-related situational variables which, we suggested, determine the degree of contextuality: size of audience, and difference in backgrounds. Her thesis that women feel more comfortable in private situations, and prefer to use a style of language specifically adapted to those situations (sometimes inappropriately when the situation is of the public type) supports our observations on the relations between contextuality, situation and gender.

It is interesting to speculate about the causes of these different communicative styles. Although there are obvious cultural influences on the way men and women communicate, recently a consensus seems to have emerged about the existence of deeper, biological differences between men and women that affect their language and thinking (Kimura, 1992). On average women are significantly better at tasks involving fluency in language, memorization of concrete items, and rote calculation. Men, on the other hand, perform better with problems requiring spatial insight and abstract, mathematical reasoning. Anastasi summarizes the effect of these biological differences in cognitive development:

> girls' acceleration in verbal communication, considered together with boys' greater ability to move about and to manipulate objects, may provide a clue to subsequent sex differences in problem-solving approaches. From early childhood, girls may learn to meet problems through social communication, while boys may learn to meet problems by spatial exploration and independent action (Anastasi, 1985: 22).

This confirms Tannen's (1992) observation that women use language preferentially for establishing social ties, while men use language preferentially for individual problem-solving. She illustrates the difference in approach with the classic situation where a couple are arguing about how to find their way in an unknown city: while the woman wants to ask directions to a passer-by, the man prefers to orient himself by studying a map.

These differences might be explained by considering the evolution of early hominids, where there would have been a clear division between male and female roles (Kimura, 1992): men would have concentrated on hunting and scavenging, which requires exploration and movement over large distances; women would have stayed more in the vicinity of their camp, gathering fruit and tubers, and caring for the children, which requires sensitivity for small details, and strong social and communicative competence. The general picture that seems to emerge is that women would be more sensitive to the immediate social and physical context, whereas men would tend to see problems more from a distance, with less attention to details, but more eye for abstract or general features.

Most of this is still speculation, but we hope that the measurement of differences in formality between male and female language may help to clarify these issues. For example, it might be used to determine to what degree the relative preferences of men for more formal expressions is dependent on culture or education.

4.3. Introversion

In personality psychology, a consensus has emerged that the most important differences in personality can be reduced to combinations of 5 basic dimensions: the big five (Digman, 1990). These were derived by several independent factor analyses of very large numbers of personality variables. The most important of these is the factor introversion/extraversion. Intuitively, extraverts are characterized as outgoing, gregarious and fun-loving, whereas introverts are seen as more quiet, reserved and pensive.

To this intuitive distinction between types of social behavior, Eysenck (1981) has added a biological dimension. According to Eysenck's theory, which has been confirmed by a number of experimental findings (Strelau, 1984), introverts are characterized by a higher level of intrinsic activation or arousal in the brain cortex. As any individual operates ideally with a moderate level of cortical arousal, the more extraverted will be inclined to look for external stimulation to reach an optimal level, whereas the more introverted people would rather try to avoid strong stimuli in order not to raise their activation level too much. This means that typical introverts are highly sensitive, reacting strongly to relatively mild stimulation, whereas typical extraverts are excitement-seekers, with a much higher endurance for loud noise, strong light, and other forms of external stress.

Extraverts and introverts also seem to have different reminiscence capabilities (Eysenck, 1971). Reminiscence is due to consolidation of the memory trace. This consolidation, which is a direct function of cortical arousal, proved to be stronger in the introverts, at least in the long run (after more than 30 minutes). Extraverts, on the other hand, showed better memory and greater reminiscence in the short run (Howarth and Eysenck, 1965; Helode, 1985).

Furnham (1990), reviewing the literature on language and personality (for native English speech), estimates that introverted speakers are likely to use a more formal style, characterized by a higher proportion of nouns, adjectives and prepositions, and a lower proportion of pronouns, verbs and adverbs. Our studies on French interlanguage referred to earlier (Dewaele, 1996a, 1998; Dewaele & Furnham, 1999, 2000) provides a few more details. In the examination situation, the degree of extraversion was found to have a significant negative correlation with the explicitness factor measuring formality. Weaker correlations were found for the informal situation and for the essays.

A possible interpretation of these results is that introverts would spend more time reflecting before they speak, whereas extraverts would be quicker to react, avoiding pauses in the conversation. Eysenck (1971) notes the introvert is more thoughtful than the extravert, taking more heed of the maxim that one should be sure brain is engaged before putting mouth into gear (p. 213). This would follow from the extraverts' need for the recurrent stimulation that a conversational interaction provides, and the introverts' preference for undisturbed, inner reflection. The longer time spent on reflection would make the introvert's speech more precise and richer in distinctions, but less fluent and less reactive to the immediate context of the conversation. This also fits in with the introverts' better long term memory allowing

them to retrieve more accurate descriptions, while the extraverts' better short term memory allows them to react and speak more quickly. This intrinsic difference in styles will be reinforced by the differential reactions of introverts and extraverts to external stress. The more sensitive introverts will become markedly less fluent in stressful situations, which interfere with their interior processes. The stress will also make them more anxious so that they become even more motivated to avoid misunderstandings (Dewaele & Furnham 2000). This may explain why the difference in formality scores was much greater in the intrinsically stressful examination situation.

4.4. Level of education

Normally, we could expect that the higher the academic level a person has reached, the richer his or her vocabulary and the wider his or her outlook. This would lead academically educated persons to express their thoughts in a more precise and less subjective way, that is to say with more formality. More generally, since the major obstacle to the use of formal descriptions is the increased cognitive load, we would expect cognitively more skilled individuals to be less inclined to avoid formality. Thus, we might hypothesize that formality would correlate positively with the general factor of intellect (also called openness to experience), which is also part of the big five(Digman, 1990).

The empirical evidence we found for this hypothesis is as yet limited. In the Dutch frequency dictionary of Uit den Boogaert (1975), word frequencies for speech of people with an academic degree are contrasted with frequencies for speech of people without such a degree (table 1). The resulting formality scores are 44 and 40 respectively. The other Dutch frequency dictionary (de Jong, 1979) compares the speech of people from a high social background with the speech of people from a low background, where background is determined on the basis of education level and occupation. The formality scores (46 and 43 respectively) differ 3 points, which is comparable to the 3 points difference between male and female speech we calculated on the basis of the same dictionary, and our more detailed analysis of these Dutch data (Dewaele 2000) similarly confirms their significance.

Interestingly, we also found that background interacts with situation: the difference in contextuality between formal and informal situations is much larger for high background people than for low background ones. A possible interpretation is that when the situation requires more formality, people with a higher education are capable of a greater shift to such a cognitively more demanding communication style, confirming our hypothesis that cognitive load is an important constraint on the degree of formality.

For written documents, our data show that more intellectual sources (scientific and technical documents, essays, broadsheet newspapers, academic prose), addressed to a more high-brow audience, are markedly more formal than sources addressed to a more average audience (family magazines, novels, fiction) (cf. tables 1 and 2).

In conclusion, we have proposed three personality variables that correlate with formality: gender, introversion and level of education Although the empirical evidence is limited, and the theoretical justification is tentative, the existence of these relations seems to match intuitive expectations. The effect of each separate variable is not that

strong (of the order of 3 or 4 points on the F-score), but it might be made more visible by combining the extreme values of the three variables. Thus, the prototypical producer of formal speech would be a male, introverted academic. The most likely person to speak in a highly contextual way would be an extraverted woman without formal education.[4]

5. Summary and conclusion

The main aim of this paper was to introduce a fundamental dimension of linguistic communication: the formality/contextuality continuum. Formal expression tries to put as much information as possible in the message itself, whereas contextual expression implicitly relies on the context shared by sender and receiver to convey part of the information. A formal style of expression is characterized by detachment, precision, and objectivity, but also rigidity and cognitive load; a contextual style is much lighter in form, more flexible and involved, but correspondingly more subjective, less accurate and less informative.

We have proposed an empirical measure for this dimension, which is based on the average degree of deixis for the most important word classes. Nouns, adjectives, articles and prepositions are used basically for context-independent expression. Pronouns, adverbs, verbs and interjections are used more frequently in contextual language. These properties were summarized by introducing an F-score for formality, in which the frequencies of the former word categories are added, the frequencies of the latter categories subtracted, and the result is normalized, so that it would vary between 0 and 100%. It was shown that this measure, though coarse-grained, reliably distinguishes more from less contextual genres of language production, for some available corpora in Dutch, French, Italian and English.

A review of several factor analyses showed that a factor similar to the F-score automatically emerges as the most important one when different samples are compared, and this in the most diverse languages. This confirms our assumption that formality/contextuality is the most fundamental and most universal dimension of stylistic variation. Given the simplicity, generality and explanatory power of this concept, the most surprising observation is that no other language researchers seem to have considered a similar model. At best, some researchers have noted the strong, recurrent patterns in their data, but lacked a good theory to explain them, while others have suggested theoretical concepts such as explicitness or indexicality, but without operationalizing them so that they could be applied to empirical data.

---

[4]The first category might be exemplified by a professor of mathematics or theoretical physics, for example Albert Einstein, and the second one by a singer or actress, say Marilyn Monroe. We leave it as an exercise for the reader to calculate the formality score of two typical expressions characterizing these well-known figures: “energy is equal to the product of mass with the square of the velocity of light”, and “I wanna be loved by you, by you, nobody else but you...”.

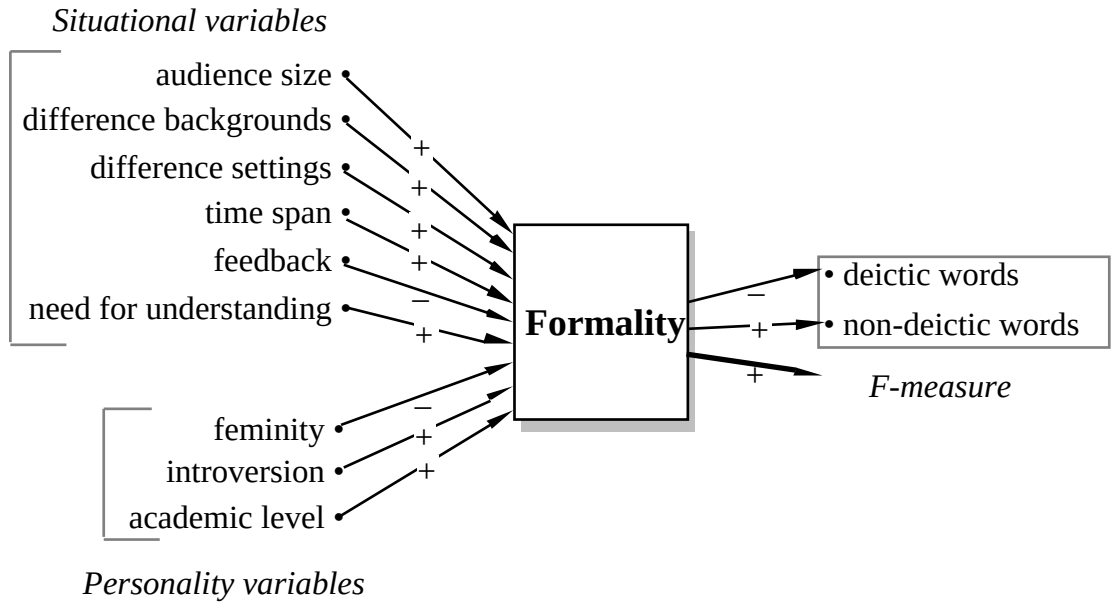

Figure 1: Summary of the formality model. Arrows with + signs denote positive correlations, – signs denote negative correlations; to the left (with arrows entering formality) are the behavioral variables that affect the formality of linguistic expressions, to the right (outgoing arrows) are the linguistic variables affected by formality; at the bottom are the abstract features by which formality is defined.

Both our theoretical model and the empirical data suggest a number of clear correlations between formality and different situational and personality variables (see Fig. 1). The formality of the language produced in a situation will increase with the importance of avoiding misinterpretation and the lack of feedback. It will decrease with the size of the shared context. This size is larger when the interlocutors are more similar or know each other more intimately, when the audience is smaller, when the sender and receiver are in the same settings, when the time interval between sending and receiving is smaller, and when a shared context has been created by previous discourse.

Moreover, contextuality appears to depend on different characteristics of the language producer. Speech is likely to be less contextual if the speaker is male, introverted and/or of a high education level. These observations can be explained by our model if we assume that: 1) women prefer involvement, whereas men prefer a more detached, independent attitude towards their conversation partner; 2) extraverts prefer on-going interaction, whereas introverts prefer undisturbed reflection; 3) people with higher education prefer precise description, whereas people without lower education prefer minimizing cognitive load.

Although none of these correlations has been fully confirmed yet, both the theoretical model and the empirical measure of contextuality we propose seem ripe for an extensive application to these and others issues in the interaction between language and situation. We hope that other researchers will adopt our formality measure and use it to test different hypotheses about language and behavior in a variety of settings.

References

AAAI-97 (1997): Fall Symposium on Context in Knowledge Representation and Natural Language, November 8-10, Cambridge, Massachusetts. The AAAI Press, Menlo Park, California.

Bar-Hillel Y. (1954) Indexical Expressions. *Mind* 63: 359-379.

Barnes B. & Law J. (1976) Whatever Should Be Done with Indexical Expressions. *Theory and Society* 3, 223-237.

Besnier, N. (1988) The Linguistic Relationships of Spoken and Written Nukulaelae. *Language* 64, 707-736.

Biber, D. & Hared, M. (1992) Dimensions of Register Variation in Somali. *Language Variation and Change*, 4, 41-75.

Biber, D. (1988) *Variation across speech and writing.* Cambridge University Press, Cambridge.

Biber, D. (1995) *Dimensions of Register Variation. A Cross-Linguistic Comparision.* Cambridge University Press, Cambridge.

Biber, D. Conrad, S. & Reppen, R. (1994) Corpus-based Approaches in Applied Linguistics. *Applied Linguistics*, 15, 2, 169-185.

Bortolini, U. Tagliavini, C. & Zampolli, A. (1971) *Lessico di frequenza della lingua italiana contemporanea*. IBM Italia.

Brown, P. & Levinson, S. (1979) Universals in language usage: Politeness phenomena, in: *Questions and politeness. Strategies in social interaction*. E.N. Goody (ed.), Cambridge University Press, Cambridge, 56-289.

Coates, Jennifer (2000). Gender differences in conversational story-telling. Paper presented at the Sociolinguistics Symposium 2000, Bristol, April 2000.

De Jong, E.D. (1979) *Spreektaal. Woordfrequenties in gesproken Nederlands*. Bohn, Scheltema & Holkema, Utrecht.

Dewaele, J.-M. (1993a) *Variation synchronique dans l'interlangue française* (unpublished PhD. thesis, Vrije Universiteit Brussel)

Dewaele, J.-M. (1993b) Extraversion et richesse lexicale dans deux styles d'interlangue française, *I.T.L., Review of Applied Linguistics* 100, 87-105.

Dewaele, J.-M. (1994) Extraversion et interlangue, in: *Profils d'apprenants, Actes du IXe Colloque international 'Acquisition d'une langue étrangère: perspectives et recherches'*, Publications de l'Université de Saint-Etienne, Saint Etienne, 173-187.

Dewaele, J.-M. (1995) Style-shifting in oral interlanguage: Quantification and definition, in: *The Current State of Interlanguage*, L. Eubank, L. Selinker & M. Sharwood Smith (eds.), John Benjamins, Amsterdam-Philadelphia, 231-238.

Dewaele, J.-M. (1996a) How to measure formality of speech ? A Model of Synchronic Variation, in: *Approaches to second language acquisition*. *Jyväskylä Cross-Language Studies* 17, K. Sajavaara & C. Fairweather (eds.), Jyväskylä, 119-133.

Dewaele, J.-M. (1996b) Variation dans la composition lexicale de styles oraux, *I.R.A.L., International Review of Applied Linguistics* XXXIV/4, 261-282.

Dewaele, J.-M. (1998a) The effect of gender on the choice of speech style, *ITL Review of Applied Linguistics*, 119-120, 1-17.

Dewaele, J.-M. (2000) *Gender, social and situational variables in the choice of speech style in native Dutch.* Paper presented at the Sociolinguistics Symposium 2000, Bristol, April 2000.

Dewaele, J.-M. (In press a) Une distinction mesurable: corpus oraux et écrits sur le continuum de la deixis. Journal of French Language Studies.

Dewaele, J.-M. (In press b) Quantifier le style dans la conversation. Une analyse de la variation sociolinguistique. Le Langage et l'Homme. Recherches pluridisciplinaires sur le langage.

Dewaele, J.-M. & Furnham, A. (1999) Extraversion: the unloved variable in applied linguistic research. Language Learning 49, 3: 509-544.

Dewaele, J.-M. & Furnham, A. (2000) Personality and Speech Production: A pilot study of second language learners. Personality and Individual Differences 28, 355-365.

Digman, J.M. (1990) Personality Structure: Emergence of the Five-factor Model. *Annual Review of Psychology*, 41: 417-440.

Fielding G. & Cooper E. (1976) Medium of Communication, Orientation to Interaction, and Conversational Style. Paper Presented at the Social Psychology Section Conference of the British Psychological Society.

Fielding G. & Fraser C. (1978) Language and Interpersonal Relations, in: *The Social Context of Language*, I. Markova (ed.), J.Wiley, Chichester, 217-232.

Furnham (A.) (1990) Language and Personality, in: *Handbook of Language and Social Psychology*, H. Giles & W.P. Robinson (eds.), John Wiley & Sons, Chichester: 73-95.

Givón, T. Function, structure and language acquisition, in: *The crosslinguistic study of language acquisition: Vol. 1*, D.I. Slobin (ed.), Hillsdale, Lawrence Erlbaum, 1008-1025.

Gorfein, D.S. (ed) (1989) *Resolving Semantic Ambiguity*. Springer Verlag, New York.

Grice, H.P. (1975) Logic and Conversation, in: *Syntax and Semantics: Vol. 9. Pragmatics*, I.P. Cole & J.L. Morgan (eds.), Academic Press, New York.

Gudykunst, W. and S. Ting-Toomey (1988) *Culture and interpersonal communication*, Newbury Park, CA: Sage.

Hall, E. T. 1976. *Beyond Culture*. Anchor Press, New York.

Halliday, M.A.K. (1985) *Spoken and written language*. Oxford: Oxford University Press.

Helode, R. D. (1985) Verbal Learning and Personality Dimensions. *Psycho-Lingua*, 15, 2: 103-112.

Heylighen, F. (1993) Selection Criteria for the Evolution of Knowledge, in: *Proc. 13th Int. Congress on Cybernetics* (Association Internat. de Cybernétique, Namur)

Heylighen F. (1999): Advantages and limitations of formal expression, *Foundations of Science*, 4:1, p. 25-56.

Hogg M.A. (1985) Masculine and feminine speech in dyads and groups: a study of speech style and gender salience, Journal of Language and Social Psychology 4. 2: 99-112.

Howarth, E. & Eysenck, H.J. (1965) Extraversion, arousal, and paired-associates recall. *Journal of Experimental Research in Personality*, 3: 114-116.

Hudson, R. (1994) About 37% of word-tokens are nouns, *Language* 70, 331-339.

Juilland, A. & Traversa, V. (1973) *Frequency Dictionary of Italian Words*. Mouton, The Hague.

Kim, Y-J. and Biber, D. 1995. A Corpus-Based Analysis of Register Variation in Korean. *Sociolinguistic Perspectives on Register Variation* . D. Bier & E. Finegan (eds.) Oxford University Press, Oxford, 157-181.

Kimura D. (1992) Sex Differences in the Brain, *Scientific American* vol. 267, no. 3 (Sept. 1992): 80-87.

Kleiber, G. (1991) Sur les emplois anaphoriques et situationnels de l'article défini et de l'adjectif démonstratif, in: *Linguistique théorique et synchronique. Actes du XVIIIe Congrès International de linguistique et de philologie romanes*, D. Kremer (ed.), Niemeyer, Tübingen, 294-307.

Leckie-Tarry, H. (1995) *Language and context. A functional linguistic theory of register*. (edited by David Birch), London-New York: Pinter.

Levelt ,W.J.M. (1989) *Speaking. From intention to articulation*, MIT Press, Cambridge, Mass.

Mazzie, C.A. (1987) An Experimental Investigation of the Determinants of Implicitness in Spoken and Written Discourse, *Discourse Processes* 10, 31-42.

Moscovici S. & Plon M. (1966) Les situations-colloques: observations théoriques et expérimentales, *Bulletin de psychologie* ,247, 702-722.

Prince, E.F. (1981) Toward a Taxonomy of given/new information, in: *Radical Pragmatics*, P. Cole (ed.), Academic, New York.

Strelau, J. (1984) Temperament and Personality. In H. Bonarius, G. Van Heck & N. Smid (Eds.), *Personality Psychology in Europe. Theoretical and Empirical Developments* (pp. 303-315) Lisse (NL) Swets & Zeitlinger.

Tannen D. (1992) You just don't understand. Women and Men in Conversation, Virago Press, London.

Tannen D. (1993) *Gender and Conversational Interaction.* Oxford: Oxford University Press

Thorne (B.), Kramarae (C.) & Henley (N.) (1983) Language, gender and society: Opening a second decade of research, in: *Language, gender and society*, B. Thorne, C. Kramarae & N. Henley (eds.), Newbury House, Rowley, MA.

Uit Den Boogaert, P.C. (1975) Woordfrekwenties. In geschreven en gesproken Nederlands. Oosthoek, Scheltema & Holkema, Utrecht.

Zampolli, A. (1977) Statistique linguistique et dépouillements automatiques, in: *Lexicologie,* Van Sterkenburgh P.J.G. (ed.), Wolters-Noordhoff, Groningen, 325-358.