Internal Report, Center "Leo Apostel", Free University of Brussels, 1999 (© Heylighen & Dewaele, 1999)

Formality of Language: definition, measurement and behavioral determinants

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. A new concept of formality of linguistic expressions is introduced and argued to be the most important dimension of variation between styles or registers. Formality is subdivided into "deep" formality and "surface" formality. Deep formality is defined as avoidance of ambiguity by minimizing the context-dependence and fuzziness of expressions. This is achieved by explicit and precise description of the elements of the context needed to disambiguate the expression. A formal style is characterized by detachment, accuracy, rigidity and heaviness; an informal style is more flexible, direct, implicit, and involved, but less informative. An empirical measure of formality, the F-score, is proposed, based on the frequencies of different word classes in the corpus. Nouns, adjectives, articles and prepositions are more frequent in formal styles; pronouns, adverbs, verbs and interjections are more frequent in informal styles. It is shown that this measure, though coarse-grained, adequately distinguishes more from less formal genres of language production, for some available corpora in Dutch, French, Italian, and English. A factor similar to the F-score automatically emerges as the most important one from factor analyses applied to extensive data in 7 different languages. Different situational and personality factors are examined which determine the degree of formality in linguistic expression. It is proposed that formality becomes larger when the distance in space, time or background between the interlocutors increases, and when the speaker is male, introverted or academically educated. Some empirical evidence and a preliminary theoretical explanation for these propositions is discussed.

Short Abstract: The concept of "deep" formality is proposed as the most important dimension of variation between language registers or styles. It is defined as avoidance of ambiguity by minimizing the context-dependence and fuzziness of expressions. An empirical measure, the F-score, is proposed, based on the frequencies of different word classes. 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 formality increases when the distance in space,

time or background between the interlocutors increases, and when the speaker is male, introverted or academically educated.

Keywords: formality, language, stylistic variation, context-dependence, fuzziness, word frequencies, personality, situation.

1. DISTINGUISHING FORMAL AND INFORMAL STYLES

A classic issue in the study of language is the measurement of variation between different genres or registers. As Labov (1972) noted, "the most immediate problem to be solved in the attack on sociolinguistic structure is the quantification of the dimension of style" (1972: 245). Stylistic variation results from the fact that different people express themselves in different ways, and that the same person may express the same idea quite differently when addressing different audiences, using different modalities, or tackling different tasks (cf. Bell, 1984, 1997). The number of possible variations is so large, though, that Labov's problem seems unsolvable as a whole.

The problem may be substantially simplified by focusing on just one aspect or dimension of style. Perhaps the most frequently mentioned of these aspects is *formality.* Everybody makes at least an intuitive distinction between formal and informal manners of expression. A prototype of formal language might be the sentence read out by a judge at the end of a trial. Prototypical informal speech would be produced in a relaxed conversation among close friends or family members. But a clear and general definition of "formality" is not obvious.

The Dictionary of Language Teaching and Applied Linguistics (Richards, Platt and Platt, 1997: 144) defines "formal speech" as follows: "the type of speech used in situations when the speaker is very careful about pronunciation and choice of words and sentence structure. This type of speech may be used, for example, at official functions, and in debates and ceremonies". This definition gives us an idea of what a formal *situation* is, but does not define formal speech as such; it just offers a hypothesis of what a speaker pays attention to in certain situations. The main criterion for formality in speech is thus non-linguistic. In a similar vein, according to Labov (1972) and Tarone (1988), the presence of channel cues: "modulations of the voice production which affect speech as a whole" (1972: 95)[1] would indicate an informal style, but, again, these characteristics reveal nothing about the intrinsic structure of (in)formal language.

Some linguists (e.g. Kirk 1988; Gelas 1988; Blanche Benveniste 1991) have tried to determine the formality level of a speech extract by considering the frequency of words and grammatical forms that are viewed as either "familiar" or "careful", such as "vous" vs. "tu" or the omission of the negative particle in sentence negations in French, and the frequency of the auxiliary "be" in English. Such a way of defining formality is, however, *ad hoc*, intrinsically limited and too dependent on the specific language and culture.

The ambiguity that surrounds the definition of formality has puzzled researchers in other disciplines. For example, Irvine (1979: 775), an anthropologist, notes that "when

---

[1] (volume, speech rate, pitch, rhythm, presence of laughter in the speech extract)

formality is conceived as an aspect of social situations, it is common to extend the term to linguistic varieties used in such situations, regardless of what those varieties happen to be like otherwise". She concludes that formality is a cover term "so general that it is not very useful as an analytic tool" (1979: 786). The lack of a good definition of formality and the quantification of the dimension of style has hampered sociolinguistic research as Labov (1972) had foreseen. Rickford & McNair-Knox (1995) point out that the decline of interspeaker or stylistic variation as a focus of research in quantitative sociolinguistics was precisely due to "the fact that investigators found it difficult to separate 'careful' from 'casual' speech in reliable and objective ways" (Rickford & McNair-Knox 1995: 265).

The underlying assumption of most approaches is that formal language is characterized by some special "attention to form" (Labov 1972), where the formal speaker tries to approximate as closely as possible the standard form and pronunciation of the language, perhaps the way it is defined in textbooks. But we should first ask why someone would want to invest more than the usual amount of attention in the form of his or her expressions.

Though we certainly can imagine occasions, such as ceremonies, rituals or examinations, where form appears important for form's sake, the most fundamental purpose of language production is still *communication*: making oneself understood by someone else. Even language that seems to have a purely social, "non-informational" function (e.g. expressing conformity to the group norm) still communicates the elementary message "I do/don't belong to the same group as you", and tries to do that as clearly as possible. We assume that language production will in general obey Grice's (1975) maxims of conversation, which include requirements of informativeness, truth, relevance, and the avoidance of obscurity and ambiguity.

In that perspective, speakers would pay more than the normal attention to form, if they would want to make sure that their expressions are not misunderstood. That would be necessary in those situations where effective communication is for some reason more difficult or more important than in ordinary circumstances. The prototypical examples we noted earlier seem to confirm this intuition: in the court situation, it is essential that no part of the verdict be misinterpreted; in the informal talk among friends, on the other hand, precise understanding is neither difficult to achieve nor very important.

This analysis leads us to distinguish two types of formality. The first one, which may be called *surface formality*, is characterized by attention to form for the sake of convention or form itself. It corresponds to the definition of the word "formal" as "rigorously observant of forms; precise, prim in attire, ceremonious" (Oxford Dictionary 1989). However, the same dictionary also lists another sense for "formal": "explicit and definite, as opposed to what is matter of tacit understanding". That second sense of the word corresponds to what we might call *deep formality*, that is, attention to form for the sake of unequivocal understanding of the precise meaning of the expression.

In the present paper we will focus on "deep" formality, because we believe that it is theoretically more fundamental, and has wider practical applications than the surface variant. In fact, we hypothesize that attention to form on the surface level will in most cases merely reflect attention to unequivocal expression on the deep level. The relatively few instances of surface formality where meaning or understanding is neglected for

decorum, thus flouting the conversational maxims (Grice 1975), could be viewed as parodies or corruptions of deep formality, which retain some stylistic attributes from their deeper origin but without the original purpose. They may be the result of ill intentions (e.g. a politician may use a formal style of language in order to create the impression that he presents precise, objective information, while he really wants to hide the exact details of his policy), or simply of rigidified conventions or traditions, where the maintenance of the initial form has taken precedence over the maintenance of the original message.[2]

Another advantage of moving the analysis to the deep level is that the structures we will find there will be more universal, and less language-specific or culture-dependent than their surface counterparts, such as the omission of the negative particle "ne" in informal oral French. Though the "deep" definition we will propose might seem more abstract or theoretical than these surface constructions, we will show that it can be easily operationalized. The resulting empirical measure will be shown to effectively distinguish language that is intuitively considered as "formal", from language belonging to typically "informal" styles of expression.

2. A THEORETICAL DEFINITION OF FORMALITY

2.1. Context-dependence

Our provisional characterization of deep formality as avoidance of ambiguity is closely related to the meaning of the word "formal" in mathematics and logic. It is a commonplace that natural languages, like English, are very different from mathematical formalisms, such as propositional calculus, in spite of apparently shared terms or concepts (e.g. "not", "and", "if...then", etc.). 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 framework and context (including the general rules or "maxims" of conversation), expressions which appear ambiguous or non-sensical when interpreted

---

[2] In the latter case, the literal meaning of the expressions has often been lost or become ambiguous. This is typical for different rituals or ceremonies where language has become "symbolic" or "poetic", that is to say open for personal interpretations. The connotative or metameaning, confirming the identity, coherence and stability of the group or tradition, however, may remain quite unequivocal. In that sense, though the formalism of rituals and ceremonies may seem purely of the surface type, there is often a deeply formal message to be found in a second order interpretation (e.g. "We all belong together, and will distance ourselves from those who don't belong"). We will not further discuss this situation as it is much more complex and less common than the general case.

separately become quite clear and logical. Grice adds that sometimes people deliberately transgress one specific rule in order to create special, "dramatic" effects, such as irony, hyperbole or metaphor. However, assuming that the person still follows the other rules, the apparent irrationality can be resolved and the expression becomes meaningful again, albeit in a more indirect, second-order way.

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", "them", which must be connected to a particular person, "here", "over there", "upstairs" which must be attached to a particular place, and "before", "now", "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, 1991, 1992). In philosophy, such expressions are usually called "indexical" (Bar-Hillel, 1954; Barnes & Law, 1976). The term "context-dependence" 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, such as anaphora. 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*.

We have provisionally characterized formality as an attempt to avoid ambiguity. We can be more specific now, and note that 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 an informal expression of the same meaning.

For example, the context-dependent expression "I'll see him tomorrow" can be rephrased more formally as "Karen Jones will see John Smith on October 13, 1999". 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, 1999, 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 context-dependent 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) maxims of quantity: the message should be as informative as is required, but not more.

2.2. Fuzziness

We must note that there are types of ambiguity which cannot be resolved by including contextual information. Sometimes the necessary information simply is not available. If no thermometer can be found, the most precise description of the temperature may be "it is hot". But does that mean that it is 25° C, or 40° C, or somewhere in between? As another example, everybody knows that "being in love" is singularly difficult to ascertain, and "am I really in love?" is one of the most often heard questions when discussing affairs of the heart. Apparently, the meaning of the word "love" is vague or fuzzy: it is difficult to distinguish instances of "love" from instances of mere "liking", "friendship", "attraction" or "infatuation".

We will reserve the term *fuzziness* to describe the situation where the reference of an expression is not unambiguously determined, even when the complete context is given. At most, in the case of a fuzzy expression some kind of quantity or probability can be established, measuring the likeliness that a particular phenomenon would be considered to belong to the class denoted by an expression. A person measuring 7 feet would thus be considered "tall" with a much higher probability than a person measuring a mere 6 feet. This is elaborated in the mathematics of fuzzy set theory or fuzzy logic (Klir & Folger 1987; Zadeh 1965).

We must note that expressions can be both fuzzy and context-dependent (cf. Ezhkova 1993). For example, a "tall" building means something different in the context of the New York skyline, than in the context of a country-side village. Similarly, the word "here" is contextual, as it will denote different places when used by people in different locations. But even if we know the exact context in which the word is used, there remains fuzziness as to the boundary distinguishing "here" from "there". In practice, it is difficult to clearly separate fuzziness and context-dependence. Both types of ambiguity need additional information to be resolved, but in the context-dependent case these data are readily available, while in the fuzzy case some effort will need to be spent in order to get the data (e.g. by measuring or more careful observation), or the

data will simply remain out of reach (e.g. even with the best instruments you cannot measure precisely how many grains of sand are to be found on the beach).

From our characterization of deep formality as minimization of ambiguity, it follows that formal styles will tend to avoid not only context-dependent expressions, but also fuzzy ones. But fuzziness cannot be eliminated without additional observation. A formal communicator would be more motivated to do that supplementary effort, whereas an informal communicator might be satisfied with a fuzzy description, since the context which is being described tends to be available for inspection anyway.

For example, imagine describing the contents of a room, first, to someone sitting with you inside that room, second, to someone you are conversing with over the telephone. In the first case, you might say "The big thing in the corner dates from the 18th century", and it would be obvious to your interlocutor what you are talking about, in spite of the context-dependence and fuzziness of the expression "the big thing". In the second case, you would have to be much more precise, stating for example "In the right corner, next to the entrance, stands a 2 meter high wooden cupboard with gold inlays, that dates from the 18th century".

In practice, formal speakers will tend to choose the least fuzzy expressions that can be applied without too much effort. But since the information necessary to resolve fuzziness is by definition not completely under the control of the communicator, while the information specifying the context is, we should expect much more variation between formal and informal styles on the level of contextuality than on the level of fuzziness.

Though we have argued that fuzziness and context-dependence will in general covary, this is not necessarily the case. It is possible to imagine expressions characterized by high fuzziness and low context-dependence, e.g. the evasive answer given by a politician to a journalist, which is meant to project an image of seriousness and objectivity, while minimizing the amount of concrete information. Similarly, we could see poetry, characterized by both subjectivity or personal involvement and very detailed description, as minimally fuzzy and maximally context-dependent. If formality is defined as a linear combination of precision (the inverse of fuzziness) and context-independence, then we could define the orthogonal dimension of *expressivity* as a linear combination of precision and context-dependence (see Fig. 1). Poetry would be highly expressive, while the politician's talk would be very low in expressivity. These rather uncommon examples illustrate our point that most language variation is to be expected along the formality dimension. Variation along the expressivity axis is less natural in the sense that it will always to some degree flout Grice's (1975) maxims of informativeness and avoidance of ambiguity, in the case of poetry in order to create unique artistic effects, in the case of the politician beating around the bush in order to simply avoid communication. This encompassing view of formality, expressivity, fuzziness and context-dependence is summarized in Fig. 1. The present paper, however, will further only discuss the formality dimension.

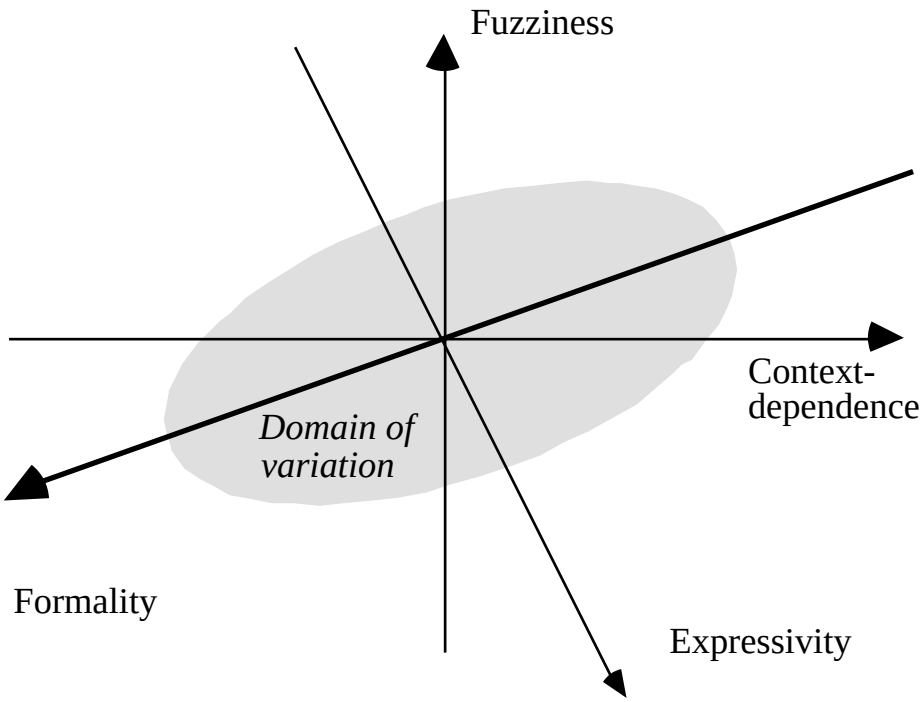

Fig. 1: Formality and expressivity as combinations of context-dependence and fuzziness. The greyed-out area represents the domain in which most variation is to be expected.

2.3. The continuum of formality

We now have come to the point where we can propose a definition of formality in the "deep" sense of avoidance of ambiguity. An expression is formal when it is context-independent and precise (i.e. non-fuzzy), that is, it represents a clear distinction which is invariant under changes of context (Heylighen 1993).

An advantage of the present definition is that it is more or less equivalent with the sense of "formal" as it is used in mathematics and the sciences. A scientific theory is called "formal" when it is expressed in a form (usually mathematical) such that there is no ambiguity as to the meaning and implications of its expressions. This implies that the same statement read by two different scientists, at different moments and in different parts of the world is supposed to be interpreted in exactly the same way. Even computers, which are totally unaware of context, should be able to interpret a fully formalized statement (Heylighen, 1991). Striving to "formalize" theories or hypotheses is an essential part of the quest for objectivity, universality and repeatability that characterizes scientific research.

It must be noted, though, that complete formal description is in principle impossible (Heylighen 1992; Van Brakel 1992). Even in pure mathematics it is recognized (through the theorem of Gödel) that it is in general impossible to explicitly state all the necessary and sufficient conditions for a particular expression to be valid. There always remains an element of indeterminacy, and completely unambiguous description is impossible. This is confirmed in the physical sciences by Heisenberg's "Uncertainty Principle", which is related to the "Observer's Paradox" in the social sciences. These different epistemological restrictions are expressed most generally by the "Linguistic Complementarity Principle" (Löfgren, 1991), which states that *no language can fully*

describe its own interpretation processes (and therefore the meaning of its expressions). On a more intuitive level, the principle can be explained by noting that the meaning of an expression can only be fixed by means of a *definition*, which explicitly states the background knowledge or information about the context needed to understand the expression. However, the definition itself contains new expressions which need to be defined themselves. But those second-order definitions again contain new terms which must be defined, ..., and so on, in an endless chase for a complete description of the world (Heylighen 1992).

On the other hand, expressions must have a minimal formality in order to be understandable at all. If the meaning changed with the slightest variation of context between the utterance of the expression and its interpretation, communication would be impossible, as the sender and the receiver of the message will never share exactly the same context. For example, there will always be a certain lapse of time passing between the moment a sender forms an expression in his or her mind, and the moment the receiver has processed that expression. Sender and receiver will also always have a somewhat different background knowledge and awareness of the present circumstances. So, a minimal invariance of meaning over changes of context is necessary.

Similarly, complete fuzziness merely signifies that any interpretation is as likely as any other one, and that implies that the expression is totally devoid of meaning or information.

We must conclude that formality is a relational concept: an expression can be more or less formal relative to another expression, implying an ordering of expressions, but no expression can be absolutely formal or absolutely informal. All linguistic expressions will be situated somewhere in between these two extremes. Where exactly on that continuum the expression will lie, depends on the choices made by the one who produces the expression, which in turn depends on the situation and the personality of the sender, as we will discuss in section 4.

2.4. Advantages and disadvantages of formality

Presently, we will just summarize the main reasons why someone would prefer formal expressions to contextual ones, or vice-versa. 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 should thus expect written language in general to be more formal than spoken language. 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).

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 ones found in dictionaries. Informal 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 context-dependent 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. Informal 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".)

By distancing itself from the immediate context, formal speech will also be less *direct* than informal one, which can make use of the salient features of the context in order to express meanings. Informal 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 formality of a speech-style will depend on the requirements of the situation, but that there will still be a subjective element, depending on whether the sender prefers accuracy over immediacy, detachment over involvement, or fears possible misinterpretation more than additional cognitive load. The most reliable way of establishing these dependencies is by empirical observation, where expressions produced in different situations or by different subjects are compared as to their overall formality, in the hope of finding recurrent relationships. In order to research such dependencies, however, we must first devise an empirical measure for formality.

3. MEASURING FORMALITY

3.1. Word category frequencies and the F-measure

Although the above theoretical definition of "deep" formality 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 formal from less formal discourses. Such a measure should be both *valid*, in the sense that what it measures effectively corresponds to formality 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 informal.

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[3]. 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 extremely difficult—if not just impossible—to perform automatically.

Our basic idea 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, referring to the spatio-temporal or communicative context. 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", "OK". In logic, deictic and anaphoric words would correspond to *variables*, which do not have a fixed referent or interpretation[4].

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 formality 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 context-dependent words belong basically to the

---

[3] 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.

[4] 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.

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 non-finite 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 context-dependent 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 context-dependent, 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 using nouns (cf. Halliday 1985), while a speaker using an informal style will prefer using verbs. This increase in verb proportion in informal 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 in 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 decrease with increasing formality of speech-styles, consists of the *pronouns, verbs, adverbs,* and *interjections.* The remaining category of conjunctions has no a priori correlation with formality. 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 formality. 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), on the use of advanced French interlanguage in different situations, provides extensive data about frequencies of different word categories[5]. A corpus of 2 speech-styles and 1 written style was collected from a group of students in three situations, in increasing order of formality: 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 formality. This led to values for the F-scores of respectively 44 (informal), 54 (examination) and 56 (essay)[6].

---

[5] Ross (1977) already attempted to use word-class distribution to distinguish literary texts.

[6] 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

One might argue that the requirements of the exam situation would rather lead to surface formality than to deep formality, as a language examiner would reward attention to form more than attention to meaning. But the argument becomes less strong in the writing task, where the form requirements are the same as in the oral exam, but where the lack of feedback and shared circumstances creates a stronger need for avoiding contextual ambiguities. Still, the results seem to confirm that word frequencies are a good measure for both types of formality.

These results could be interpreted as a mere peculiarity of interlanguage or of exam situations. 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 French, we found similar variations of word frequencies between more and less formal 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 formal genres, such as scientific texts (F=66) or (serious) newspapers (F=68), lead to much higher formality scores than those from more informal 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 and Fig. 2.

---

language will tend to restrict the higher precision of expression which would otherwise be expected for written essays.

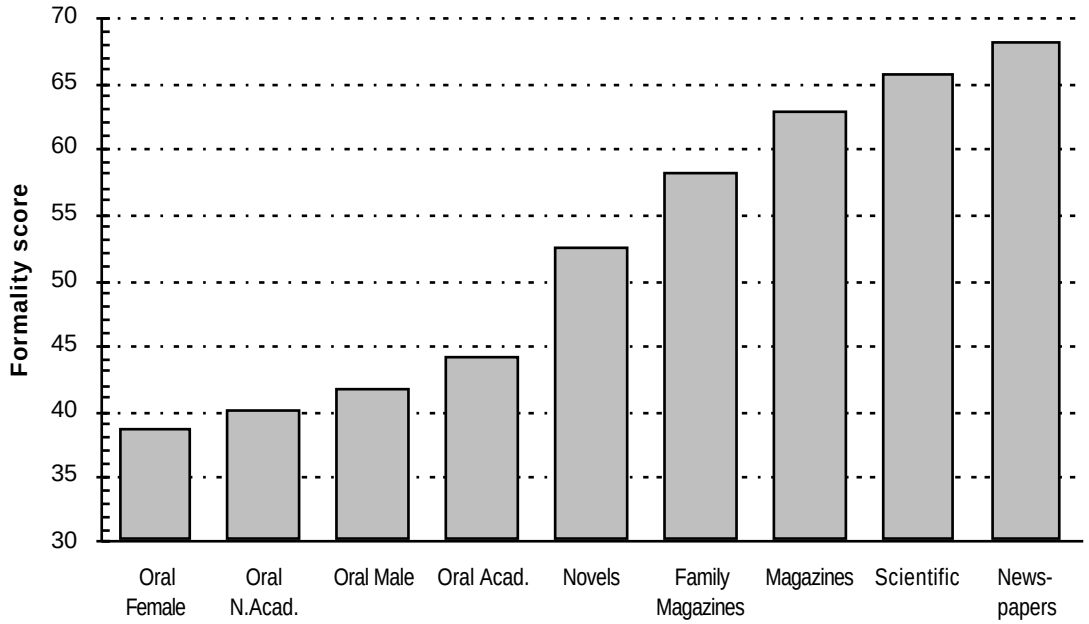

Figure 2: formality scores for Dutch language coming from different fields

	"formal" categories				*"deictic" categories*				
	Nouns	Articles	Prepos.	Adject.	Pronouns	Verbs	Adverbs	Conj.	**Forma -lity**
Oral Female	10.40	6.89	5.86	8.09	16.95	19.35	17.45	7.47	**38.7**
Oral N.Acad.	12.75	8.50	6.34	6.71	16.01	18.80	19.31	6.34	**40.1**
Oral Male	11.48	8.16	6.69	7.63	15.84	18.45	16.53	7.05	**41.6**
Oral Acad.	13.16	9.58	7.91	7.13	13.96	17.75	17.88	7.13	**44.1**
Novels	18.52	10.48	10.26	10.00	13.25	20.62	10.47	6.06	**52.5**
Fam. Magaz.	21.78	9.77	12.21	11.14	10.09	18.71	9.74	6.39	**58.2**
Magazines	24.20	11.61	13.90	10.93	8.55	17.68	8.73	4.34	**62.8**
Scientific	23.10	15.00	13.75	10.75	6.71	16.58	7.98	5.98	**65.7**
Newspapers	25.97	14.68	14.54	10.57	5.62	16.69	7.21	4.70	**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 "deictic" 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 context-dependent 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 and Figure 3).

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.

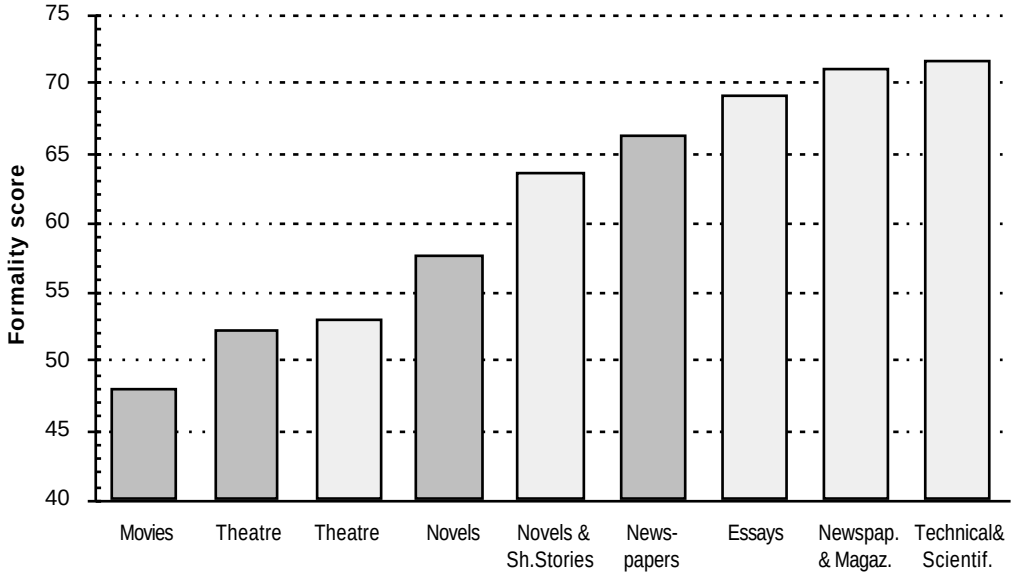

Figure 3: formality scores for Italian language coming from different fields. Darker grey columns are based on data from dictionary A, lighter grey refers to dictionary B (see text).

	"explicit" categories				*"deictic" categories*					
	Nouns	Articles	Prepos.	Adject.	Pron.	Verbs	Adverbs	Interj.	Conj.	**Formality**
Movies A	13.37	8.29	8.62	5.08	1.62	27.03	9.98	0.77	5.98	**48.0**
Theatre A	14.84	10.15	9.44	5.51	1.43	24.48	8.71	0.77	5.55	**52.3**
Theatre B	13.96	10.22	10.54	4.83	1.37	23.94	8.14	0.13	7.17	**53.0**
Novels A	16.72	13.79	14.04	5.58	8.50	20.05	6.45	0.13	6.42	**57.5**
Novels B & Sh.Stories	18.19	16.03	15.45	6.74	7.04	17.65	4.45	0.07	6.27	**63.6**
News-papers A	18.92	16.80	16.73	7.70	5.10	17.51	4.86	0.02	5.15	**66.3**
Essays B	18.95	16.91	17.15	8.09	5.75	12.90	4.15	0.03	6.95	**69.1**
Newspap. & Magaz. B	20.41	18.35	18.39	8.35	4.29	15.41	3.47	0.01	5.27	**71.2**
Technical& Scientif. B	18.63	17.99	20.17	7.56	4.27	12.73	4.12	0.00	6.00	**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 the present concept of formality 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 formality. 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 and Figure 4.

	explicit categories			deictic categories			
	Nouns	Prepos.	Adject.	Pronouns	Verbs	Adverbs	**Formality***
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.*

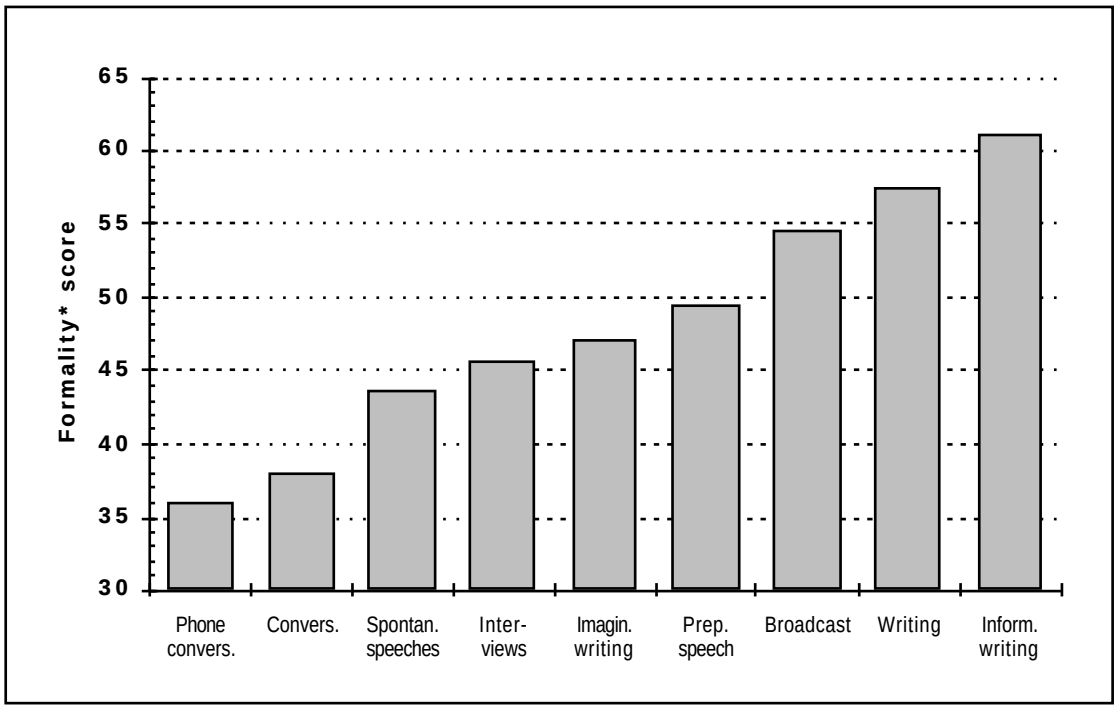

Figure 4: formality (lacking frequencies of some word categories) scores for English language coming from different fields.*

Again, we note that the formal categories mostly increase together with formality, while the deictic 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 less formal 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 less formal than the "interviews", and where boys' language is more formal 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. Related measures

Zampolli's (1977) and Hudson's (1994) cluelessness as to a theory explaining the very clear patterns in their data seems surprising. Surely, other linguists must have proposed models of similar variations. Let us review some existing proposals for stylistic dimensions related to formality.

Hasan (1984) introduces the concept of explicit and implicit style. Although she states that the degree of implicitness is determined purely quantitatively, "by comparing the proportion of explicit units to the implicit ones" (p. 110), she finally rejects quantitative grading and thus fails to present a workable empirical definition. Leckie-Tarry (1995) develops Hasan's theory and states that in an 'explicit' text the context of the situation is not immediately apparent and that "a greater load falls on the linguistic structures to convey meaning" (p. 123). According to Leckie-Tarry, words can be classified on a continuum of explicitness, and range from non-explicit core words which are not very field- or genre-specific to highly explicit words, like technical words that have strong associations with particular fields and hence particular registers (p. 127). Leckie-Tarry does not, however, propose a way of empirically measuring the degree of explicitness of a text.

Halliday's (1985) well-known "lexical density", was proposed as a measure for distinguishing written from spoken styles of language. As we argued earlier, spoken and written language will also differ markedly in formality, so that lexical density might be expected also to be a measure of formality. The measure is defined by the proportion of "content" words, which are dense in information and typical of more formal styles, to "function" words, which have a low information content and are mainly used to connect the content words together. Though related to F, lexical density differs in some essential aspects. Most verbs are considered content words just like nouns, while in our system they are classified as deictic words, separate from the nouns. Articles and prepositions, on the other hand, are function words, but in our analysis they are classified with the context-independent words. This makes lexical density less reliable as a measure or formality. Halliday (1985: 75) indirectly acknowledges this by noting that written language is not only characterized by high lexical density, but also by high nominalization, a feature marking the preference of nouns over verbs (and adjectives over adverbs) typical for styles avoiding contextuality.

Such a distinction between nominal and verbal styles is a recurrent theme, identified as fundamental by many authors, including Wells (1960) and Brown and Fraser

(1979). The general interpretation tends to be that nouns are more "static" and verbs more "dynamic", but this does not help much in elucidating why certain registers are more verbal or nominal than others.

A link with formality is noted by Brown & Levinson (1979), in their analysis of polite speech in English: "degrees of negative politeness (or at least formality) run hand in hand with the degree of nouniness. [...] formality is associated with the noun end of the continuum" (1979: 212). They explain this phenomenon by noting that "with the progressive removal of the active 'doing' part of of an expression, the less dangerous it seems to be" (1979: 213). This hypothesis seems rather *ad hoc*. Although politeness may seem more closely related to surface formality than to its deep variant, we might assume that it shares sufficient characteristics with formality as context-independence (implying avoidance of directness, involvement, and potential misunderstandings) to apply our analysis. If that is true, the "nouniness" of polite speech, and its corresponding reduction of the proportion of verbs, confirms our above analysis, which sees verb frequencies decreasing and noun frequencies increasing with an increase in formality. Unfortunately, Brown & Levinson did not provide data on the other word categories.

Fielding & Fraser (1978), in a study on language and interpersonal interaction, uncovered a similar stylistic variable, which is closely related to the F-factor:

> "The nominal-verbal factors were defined in terms of the ratios between the number of nouns and verbs, and between the number of nouns and pronouns, and by the proportions of parts of speech associated with the nominal (for example, nouns, adjectives and articles) and verbal (for example, auxiliary verbs, and adverbs) aspects of the language system" (1978: 223).

They found that:

> "the nominal style is likely to be more monotonous, less personal, and more formal. It appears to be a carefully considered and closely monitored production. The verbal style, on the other hand, is characteristic of spontaneous, unreflective speech. It is immediate, informal and varied" (1978: 226).

They further note that "this noun versus verb distinction, together with its related word classes, represents a fundamental and perhaps universal grammatical distinction" (1978: 223), without, however, offering a more profound interpretation of this distinction.

3.4. Formality as a universal factor

In spite of the empirical confirmations, our definition of F may seem to some degree arbitrary, just another one of these many related, but different, dimensions proposed by different authors, which all correlate to some degree with variations such as written vs. oral, but whose underlying motivation is debatable. We will now 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, in press a) 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". It is practically identical to formality 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 formality variable. This variation is apparently due to the personal preferences of the subjects for more or less formal 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 informal) 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 formality (or rather its opposite, context-dependence). 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 explains that "there are a larger number of positive features on dimension 1 (...) reflecting direct interaction, focus on the immediate circumstance and personal attitudes or feelings, fragmentation or reduction in form, and a less specific, generalized context" (Biber 1995: 143).

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 an informal 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 formality. 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 language of 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 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 formality appears as *the* most important and universal feature distinguishing styles, registers or genres in different languages.

3.5. Further extensions of the formality measure

The main criticism that can be raised against the present measure of formality is that it is much too coarse, reducing stylistic variations to mere frequencies of the most general word categories. Yet the measure seems to do its job, unambiguously distinguishing types of language which we would intuitively and theoretically expect to differ in formality. The advantage of such a coarse-grained approach is that it facilitates the collection and processing of data for different samples or styles.

A second advantage of working at such a high level of generality is that the resulting measure is relatively independent of language. We have shown that the measure is applicable at least to English, French, Dutch and Italian. We expect that it would be easily generalized to further languages. Even if certain word categories (say, articles or pronouns) would not exist in a particular language, we may assume that it will still be possible to distinguish more deictic from more explicit word (or morpheme) categories, in a way similar to the one we used. It would then suffice to add the frequencies of the predominantly explicit categories and to subtract the frequencies of the predominantly deictic categories in order to get an overall formality measure. We would like to stress that the resulting values for the measure can only be used for comparing excerpts within the same language.

Within a given language, it is in principle possible to refine the formality measure, taking into account more subtle differences in formality than the ones between the most general word categories. That would make the measure more precise, allowing finer distinctions between texts and a more reliable measurement of formality for small samples. At present, a sample would probably need to contain a few hundred words for the measure to be minimally reliable. For single sentences, the F-value should only be computed for purposes of illustration: there are too many syntactical, semantical and pragmatical subtleties and exceptions involved to distinguish more context-dependent from more explicit sentences by means of lexical category frequencies alone.

A simple way to refine the F-measure would consist in subdividing the abstract categories into more specific ones, for example distinguishing different types of pronouns, verbs and articles. With the resulting, larger set of variables a new factor analysis can be carried out. We have done this with the data of the corpus of advanced French interlanguage (Dewaele 1993a), starting with 27 variables denoting more fine-grained categories. Something similar to the explicitness factor still comes out first, but it explains only 22% of the variance. This could be expected, since a much larger

number of variables allows for many more sources of variation different from formality.

The correlations of the explicitness factor with the variables are similar to those with the more coarse-grained word categories, although it turns out that some subcategories show an opposite trend to the one of the global category. For example, although determiners show an overall positive correlation with explicitness or formality, the subcategory of "indefinite" determiners (e.g. "some", "certain"...) has a slight negative correlation. This could be expected, since rather than adding explicit information about the context, they make the meaning of the subsequent noun more fuzzy. The general result of the analysis seems hardly more informative than the result of the more coarse-grained analysis, though. In most circumstances it would not seem worth the additional effort.

Ideally, we could imagine a very refined measure where each word (or at least each of the most frequently used words) would get an average degree of formality. A relatively straightforward method to achieve this might consist in determining the degree of correlation between the most frequent separate words and the existing, coarse-grained measure. Positively correlating words would then be assigned to the formal category, negatively correlating ones to the contextual category. Most likely such an analysis would uncover some words behaving contrary to their general category. For example, the word "thing", being a noun, should be put in the formal category if we follow the coarse-grained procedure. It seems likely, though, that it will be effectively more common in informal discourses, thus fitting better in the deictic category.

4. Behavioral determinants of formality

As the formality 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 formality. 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 formality 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 less formal than articles or books, without possibility for reply, as confirmed by the data from Biber (1988).

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". 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.

As we argued, 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 charateristics 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 (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. We would venture that the highest degree of informality will be found among identical twins that were raised together, who completely share their cultural, social and even biological backgrounds.

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 less "informational" (formal) than professional letters. Our study of French interlanguage (Dewaele 1993a, 1996a, 1996b) 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 way 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 & Coope (1976) found that conversations over the intercom are more nominal (formal) than face-to-face conversations. Moscovi & 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 less formal score than face-to-face conversations.

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 lower formality than speeches prepared at an earlier moment (table 3). Another way to test this proposition empirically might consist in measuring the formality 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 less formal 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.

4.2. Gender

Perhaps the most visible and deeply rooted characteristic distinguishing people is their sex. 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 less formal 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). This general tendency is confirmed by our study of advanced French interlanguage (Dewaele 1996a, 1998a), where the female part of the group scored F=39 on average in the informal situation, whereas the male group scored an average F=45, an overall difference of 6 points. This difference was found to be statistically significant ($p=0.013$).

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.

More surprising is the finding that men omit the French negative particle "ne" significantly more often than women (Dewaele 1992). Such omission is usually interpreted as characteristic of a more vernacular speech (cf. Blanche-Benveniste 1991). Also in other studies it was found that women's speech tends to approximate the norms more closely than men's speech. Labov's "Principle 1" states that: "For stable sociolinguistic variables, men use a higher frequency of non-standard forms than women" (Labov 1990: 210). This would mean that women's speech is more formal in the "surface" sense, but less formal in the "deep" sense. This can be interpreted in different ways. Since formal speech demands more cognitive effort, we might expect a trade-off between content and form: as (deeply) formal speakers tend to concentrate on the informational content of their expressions, they will pay less attention to norms related to the surface form. Trudgill (1974), on the other hand, suggests that a "tough", "working-class" language is considered somehow more manly than a more refined speech

To clarify the issue, we first must interpret the apparent preference of women in informal conversations for less formality at the deep level. From socio-linguistic and psychological studies (e.g. Hogg 1985, Tannen 1993), 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 (1990, 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 context-dependence 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 (me, you, him, etc.), adverbs, inflected verbs and interjections. It also explains why the

difference in formality between men and women disappeared 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 formality: 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 formality, 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. Using functional magnetic resonance imaging, Shaywitz et al., (1995) found that language functions are more highly lateralized in males but represented in both cerebral hemispheres in females. These data provide the first clear evidence of gender-differences in the functional organization of the brain for language. The authors suggest that these differences might explain why more men suffer from language disorders like dyslexia, loss of speech and stuttering than women.

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.

Such differences may 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 independent exploration and movement over large distances. This would select for a good sense of orientation and an ability to manipulate and aim projectiles and tools. Women, on the other hand, would have stayed in the vicinity of their camp, gathering fruit and tubers, which requires a sensitivity for, and good memory of, small details. Moreover, because of pregnancy and child care, women would have been intrinsically more dependent on the group and on their relations with other individuals, including partner and children. This would have selected for strong social and linguistic 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 consider problems in a more detached way, with less attention to the subtleties of social interaction, but more eye for abstract and spatial features. This would explain women's involvement in the social context of a conversation, and the concurrent reduction of deep formality in their speech. At the same time, it may explain why women excel in surface formality, since obeying the social and linguistic norms for correct pronunciation and grammar is for women both easier and more important than for men.

Most of this remains mere speculation, but we hope that the measurement of differences in formality between male and female speech may help researchers to clarify these issues. For example, it might be used to determine to what degree the relative preference 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).

Since extraverts are more talkative and prone to start a conversation, they are usually expected to be better language learners than introverts. Language researchers were extremely disappointed when it appeared that extraversion did not correlate with language test results (Naiman et al., 1975). For two decades this finding was quoted, but never challenged, in applied linguistic studies (Dewaele, 1994). This negative publicity was so strong that many researchers (e.g. Brown 1987; Skehan 1989) seemed to believe that no significant link could be expected between extraversion and any linguistic measure.

However, when we look at language formality rather than grammatical correctness scores, a clear relationship emerges. 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, 1998a; Dewaele & Furnham, ms.) 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.

As can be expected, linguistic variables that correlate with formality also appear to correlate with introversion. Introverts' speech seems characterized by a higher lexical richness in certain situations (Dewaele 1993b, Carell, Prince & Astika 1996). On the other hand, extraverts' speech seems more fluent than that of introverts (Di Scipio 1968; Tapasak et al. 1978). One should be careful in one's definition of fluency, though, as different measures may lead to different results (cf. Muniz Fernandez & Yela Granizo 1981): measures including lexical richness would tend to benefit the introverts.

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, ms.). 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. 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 informal way would be an extraverted woman without formal education.[7]

5. SUMMARY AND CONCLUSION

[7]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...”.

We have extended the linguistic concept of formality, which can be generally characterized as "attention to the form of expressions", by subdividing it into two parts: *surface formality*, characterized by attention to form for the sake of convention, and *deep formality*, characterized by attention to form for the sake of clear understanding. We have argued that the deep part is the most important one, and suggested that the surface variant will inherit most of its stylistic features from the deep version.

We have elaborated the definition of deep formality by noting that formal language is an attempt to avoid ambiguity by minimizing the context-dependence and fuzziness of expressions. An expression is defined as context-dependent if its meaning is clear, but only to someone aware of the context in which it is produced. An expression is defined as fuzzy if its meaning is imprecise even when the context is known. Since fuzziness basically results from an intrinsic lack of information about the thing being described, a sender will have much more control over the contextuality than over the fuzziness of his or her expressions, so that contextuality may be assumed to be a better indicator of the intended degree of formality.

A formal style will be characterized by detachment, precision, and "objectivity", but also rigidity and cognitive load; an informal style will be much lighter in form, more flexible, direct, and involved, but correspondingly more subjective, less accurate and less informative.

We have proposed an empirical measure for formality 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 context-dependent 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 formal 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 genres are compared, and this in the most diverse languages. This confirms our assumption that formality 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.

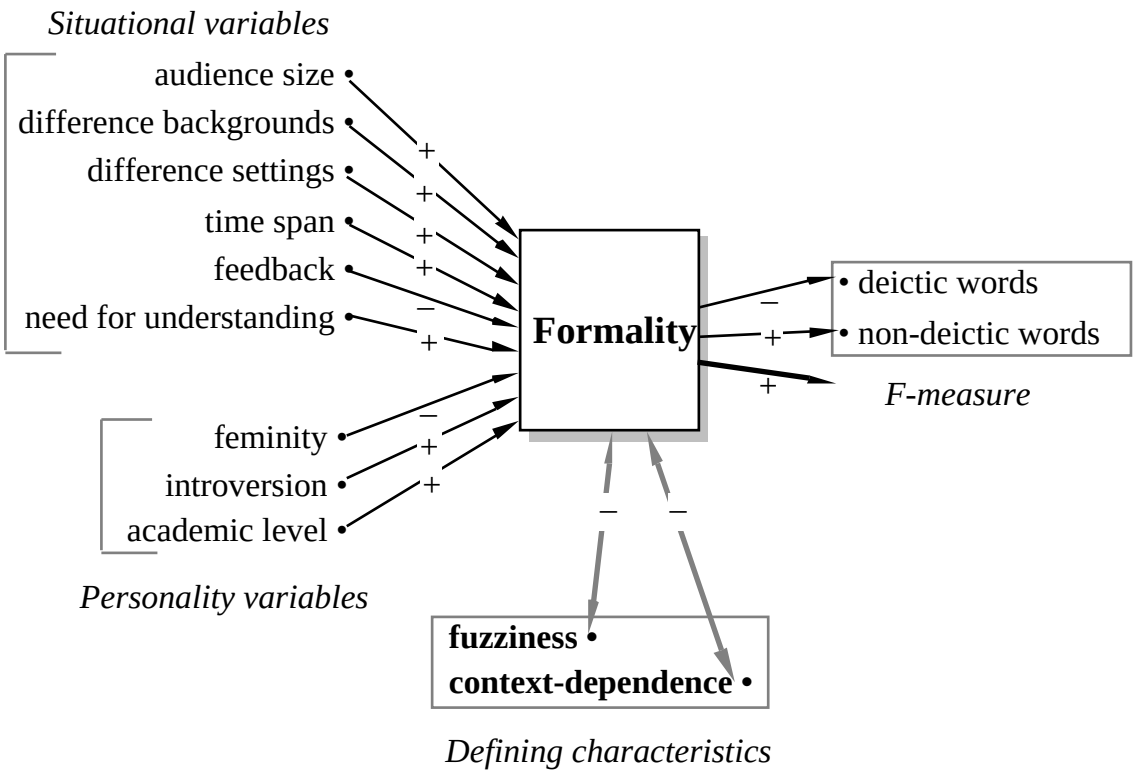

Figure 5: Summary of the formality model. Arrows with "+" signs denote positive correlations, "–" signs denote negative correlations; to the left (with arrows entering formality) are the behavioral variables that affect the formality of linguistic expressions, to the right (outgoing arrows) are the linguistic 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 behavioral variables (see Fig. 5). 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, 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, formality appears to depend on different characteristics of the language producer. Speech is likely to be more formal 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 immediate interaction, whereas introverts prefer undisturbed reflection; 3) people with higher education prefer precise description, whereas people without higher education prefer minimizing cognitive load.

Although none of these correlations has been fully confirmed yet, both the theoretical model and the empirical measure of formality we propose seem ripe for an extensive application to these and others issues in the domain of language and behavior.

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

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.

Bell A. (1984) Language Style as Audience Design. *Language in Society*, 13, 2, 145-204.

Bell A. (1997) Language Style as Audience Design in *Sociolinguistics. A Reader and Coursebook.* N. Coupland & A. Jaworski (eds.), London: Macmillan, 240-250

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.

Blanche-Benveniste, C. (1991) *Le français parlé. Etudes grammaticales.* Editions du CNRS, Paris.

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

Brown, P. & Fraser, C. (1979) Speech as a Marker of Situation, in: *Social Markers in Speech*. K.R. Scherer & H. Giles (eds.), Cambridge University Press, Cambridge, 33-62.

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.

Carell, P.L, Prince, M.S. and Astika, G.G (1996) Personality types and language learning in an EFL context, *Language Learning*, 46, 1, pp. 75-99.

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

Dewaele, J.-M. (1992) L'omission du 'ne' dans deux styles d'interlangue française, *Interface. Journal of Applied Linguistics*, 7.1: 3-17.

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. (In press a) La composition lexicale de styles oraux et écrits, *Language and Style*.25, 1 (winter 1992)

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. (1998b) Speech rate variation in 2 oral styles of advanced French interlanguage in: *Contemporary Approaches to Second Language Acquisition.* V. Regan (ed.), Dublin: University College Academic Press, 113-123.

Dewaele, J.-M. & Furnham, A. (1998) Extraversion: the unloved variable in applied linguistic research., ms.

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

Ezhkova I. V. (1993) A Contextual Approach for AI Systems Development, in: Fuzzy Logic in Artificial Intelligence: Proc. of the 8th Austrian Artificial Intelligence Conference FLAI'93, E. P. Klement and W. Slany (ed.), (Springer, Berlin), 2.

Fielding G. & Coope 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.

Gelas, N. (1988) Dialogues authetiques et dialogues romanesques, in: *Echanges sur la conversation*, Editions du CNRS, Paris, 323-333.

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.

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

Hasan, R. (1984) Ways of saying: ways of meaning. in: R. P. Fawcett, M.A.K. Halliday, S.M. Lamb, A. Makkai (eds.), *The semiotics of Culture and Language.* Vol. 1 *Language as Social Semiotic* (pp. 105-162) London & Dover: Pinter.

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

Heylighen, F. (1991) Design of a Hypermedia Interface Translating between Associative and Formal Representations, *International Journal of Man-Machine Studies* 35, 491-515.

Heylighen, F. (1992a) From Complementarity to Bootstrapping of Distinctions: a Reply to Löfgren's Comments on my Proposed 'Structural Language', *International Journal of General Systems* Vol 20, Number 4.

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

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.

Irvine, J.T. (1979) Formality and Informality in Communicative Events, *Amercian Anthropologist* 81, 773-790.

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.

Klir, G. & Folger, T. (1987) *Fuzzy Sets, Uncertainty, and Information.* Prentice Hall, Englewood Cliffs, NJ.

Labov (W.) (1972a) Contraction, Deletion and Inherent Variability of the English Copula, in: *Language in the Inner City: Studies in the Black English Vernacular*, University of Pennsylvania Press, Philadelphia.

Labov, W. (1972b) *Sociolinguistic Patterns.* University of Philadelphia Press, Philadelphia.

Labov, W. (1990) The intersection of sex and social class in the course of linguistic change, *Language Variation and Change* , 2, 205-254.

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.

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

Naiman, N.; Frohlich, M.; Stern, H.H. & Todesco, A. (1978) *The Good Language Learner.* Toronto: Ontario Institute for Studies in Education (1° edition 1975)

Oxford English Dictionary (1989) Clarendon Press, Oxford.

Paradis, M. (1994) Neurolinguistic aspects of implicit and explicit memory: Impliciations for Bilingualism, *Implicit and explicit second language learning*, N. Ellis (ed.), London, Longman, 393-419.

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

Richards, J. Platt, J. & Platt, H. (1997) *Dictionary of Language Teaching and Applied Linguistics.* London: Longman.

Rickford, J. R. & McNair-Knox, F. (1995) Adressee -and Topic- Influenced Style Shift. A Quantitative Sociolinguistic Study, in: *Sociolinguistic Perspectives on Register Variation* , D. Bier & E. Finegan (eds.), Oxford University Press, Oxford, 235-276.

Robinson, P. (1995) Task Complexity and Second Language Narrative Discourse, *LanguageLearning* 45:1, 99-140.

Ross, D. (1977) The use of word-class distribution data for stylistics: Keat's sonnets and chicken soup, *Poetics* 6, 169-196.

Shaywitz B.A., Shaywitz S.E., Pugh K.R., Constable R.T., Skudlarski P., Fulbright R.K., Bronen R.A., Fletcher J.M., Shankweiler D.P., Katz L. & Gore J.C. (1995) Sex Differences in the Functional Organization of the Brain for Language, *Nature*, 373, 6515, 607-609.

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

Tapasak, R., Roodin, P.A., Vaught, G.M. (1978) Effects of Extraversion, Anxiety, and Sex on Children's Verbal Fluency and Coding Task Performance. *The Journal of Psychology* 100, 1, 49-55.

Tarone, E. (1988) *Variation in Interlanguage*. Edward Arnold, London.

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.

Trudgill P. (1974) *The social differentiation of English in Norwich.* Cambridge: Cambridge University Press.

Tulving, E. (1985) Precis of 'Elements of Episodic and Semantic Memory', *Behavorial and Brain Sciences* 7, 223-238.

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

van Brakel, J. (1992) The Complete Description of the Frame Problem, *Psycoloquy* 3 (60) frame-problem 2.

Wells, R. (1960) Nominal and Verbal Style, in: *Style in Language*, T.A. Sebeok (ed.), MIT Press, Cambridge, Mass., 213-220.

Zadeh, L.A. (1965) Fuzzy Sets and Systems, in: *Systems Theory*.?. J. Fox (ed.), Polytechnic Press, Brooklyn NY, 29-37.

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