

# Language and the Theory of the Firm<sup>1</sup>

Jacques Crémer

Université de Toulouse, IDEI-GREMAQ and CEPR

Luis Garicano

University of Chicago and CEPR

Andrea Prat

London School of Economics and CEPR

November 24, 2005

<sup>1</sup>A previous version of this paper circulated under the title: ‘Codes in Organization.’ We thank Phillippe Aghion, Karen Bernhardt, Wouter Dessein, Matthias Dewatripont, Bob Gibbons, Jerry Green, Daniel Hoffman, Augustin Landier, Jean Tirole, workshop participants at the CEPR/Toulouse Organizations workshop, the NBER Organizational Economics Workshop, and at various universities for useful comments and Pedro Vicente for outstanding research assistance. Garicano thanks the Toulouse Network on Information Technology and the Microsoft Corporation for their financial support.

### **Abstract**

An organization will often use a specialized technical language that is understood by its members but not by the rest of the world. We develop a theory of optimal organizational languages and identify a key trade-off between facilitating internal communication and encouraging communication with other organizations. “Dialects” tend to be suboptimal: two organizations will either share the same language or develop two entirely distinct set of technical words. This endogenous discontinuity in communication structure is reflected in a discontinuity in firm structure. We explore a number of predictions of our model and we relate them to changes in firm structure associated to technological progress. Our theory reconciles two recent phenomena within organizations: the increase in information centralization and the reduction in hierarchical centralization (‘empowerment’).

# 1 Introduction

"We were late. Whether it be MPLS over ATM, whether it be precedent bit over IP."

John Bloomer, Enron employee, responding in trial to questions from Enron Task force prosecutor Ben Campbell (*Houston Chronicle*, May 11, 2005).

Agents often use technical languages to communicate with others about contingencies that are specific to their common environment. In Hollywood for example, a movie “has legs” when it will play for a sustained period of time; a star can “open a movie” when she can be counted on to ensure that the film has a strong box office. Similarly, economists say something is “Pareto optimal,” or a pair of strategies is a “Nash Equilibrium.” These “codes,” to use the term introduced by Arrow (1974), economize in communication costs, as they allow for a lot of information to be transmitted efficiently. However, such improvements in communication come at the cost of limiting communication with agents from outside the group.

In this paper, we study the tradeoffs determining the adoption of specific technical languages or “codes” by organizations and how they shape in turn different aspects of organizational design. Since agents are boundedly rational the need to learn specialized codes constrains the scope of the organization: a specialized code facilitates communication within a service or function, but limits communication between services, and thus makes coordination between them more difficult. The essential trade-off determining the scope of the language chosen is thus one between specialization and coordination. An organization that wants to capture more synergies by expanding its scope must acquire a more vague and imprecise common language to facilitate coordination among its units.

The coordination difficulties imposed by incompatible languages can have severe consequences. Consider for example an incident in the Persian Gulf between US military services on April 14, 1994. On that date, two US Airforce F-15 fighters shot down two US Army Blackhawk helicopters over the Iraq no-flight zone, killing everyone inside. The tragedy took place in broad daylight and involved highly experienced and qualified crews on both sides, who, furthermore, were monitored by a US Airforce AWACS (airborne warning and control system) flying above them.<sup>1</sup> Investigators found no individual guilty; the tragedy was the result of grave organizational dysfunctions. In particular, communication was hindered by misunderstandings resulting from the incompatibility between

---

<sup>1</sup>The account that follows is from Snook (2000).

the ‘code’ that governed Army operations with the code used by the Airforce. Three instances are particularly striking. First, the word ‘aircraft’ in the order forbidding the entry by aircraft into the No-Flight zone before the enemy radars were ‘sanitized’ by F-15s was understood by the Air Force to include helicopters, but by Army pilots to exclude helicopters (Snook, 2000:163). As a consequence, the Air Force pilots did not expect any American helicopter to be present in the no-fly zone, while the Army pilots did not think they were breaching the order. Furthermore, the AWACS crew thought it was responsible for airplanes, but not helicopters (Snook, 2000: 163). Second, the two key acronyms concerning the no-fly zones, AOR (Area of Responsibility) and TAOR (Tactical Area of Responsibility) were understood differently by Army and Air Force: ‘To the Army, AOR meant the area outside northern Iraq; to the Air-Force, it meant just the opposite’(Snook, 2000:157) These different codes translated into crucially different interpretations concerning both the ‘where’ could aircraft be and the ‘when’ they could be there, and contributed to the misperception of the two Blackhawks as enemy helicopters by the fighter pilots. Third, the Air Force and the Army helicopters interpreted differently, for a long period of time up to the accident, the rules governing the electronic exchanges used to identify other aircraft as friendly or foe (the so-called IFF system, for ‘Identify Friend or Foe’). As a result, the Army helicopters answered the Airforce pilots electronic query with the wrong code, and were identified as enemies (Snook, p. 2000:157). The Air Force pilots saw US helicopters where (they thought) they should not be, when they were not expected to be, using a wrong frequency of IFF, and shot them down as enemies.

For less beligerant examples, consider the partition of the color spectrum into discrete (yellow, red, orange) categories. Or, in a business setting, the allocation of cost and revenue items to accounting categories. In the first example, different cultures set the boundary between color at different places in the spectrum (Sperber and Hirschfeld. 1999). Similarly, different firms place boundaries between different cost categories in different places (see e.g. Herbold, 2004:58).

Our analysis proceeds in two steps. We first present a simple theory of language and characterize the properties of optimal organizational languages. Our theory builds on previous informal discussions by Arrow (1974), which emphasize the role of specialized codes in reducing the constraints on organizational performance imposed by individual bounded rationality. We then use this theory to understand the constraints that the need for specialized languages and individual bounded rationality impose on firm scope and structure, and derive testable empirical implications.

Section 2 begins with a simple model of codes. Certain agents in the organization receive a stream of heterogeneous problems (e.g. customers with certain needs, patients with certain symptoms). To solve these problems, these agents must communicate with other agents in the organization who hold specialized

knowledge (e.g. production engineers, medical specialists). Communication relies on a common code, shared by the members of the organization. As agents are boundedly rational, they can only learn a given number of words to describe the characteristics of the problems they encounter. Hence, the problem-solving capability of an organization depends on the code it uses. Following Arrow, we make a key assumption: the organization is, at least to a certain extent, able to determine the code its members adopt.<sup>2</sup>

We characterize the properties of efficient codes. Efficient codes use precise words for frequent events and vague words for more unusual ones. A more unequal distribution of events increases the value of the creation of a specialized code, since the precision of the words can be more tightly linked to the characteristics of the environment.

We show that when more than two agents communicate with one another, bounded rationality imposes sharply decreasing returns to the diversity of codes. Tailoring words to the needs of particular agents restrains the group of agents who can usefully learn them. As a consequence, homogeneously skilled agents will use either entirely separate codes or common codes: “dialects” cannot be optimal. This code commonality is a key determinant of the decreasing returns to scope in organizations, and shapes both organizational scope and their use of integrating mechanisms.

In Section 3 we study the implications of our theory of language for the organization of firms. First, our theory generates a bounded-rationality based theory of firm scope. Broadening the scope of a firm is justified by the fact that it can exploit some synergy. However, firms cannot capture the synergy and leave everything else the same. Capturing the synergies requires some coordination and communication between services, which requires in turn that they speak a common language. Since a common language cannot be as suited to the needs of the specialized individual services, the scope of the firm is limited. We also identify the variables that determine the terms of these trade-offs: a broader firm with a common code is more likely when the degree of synergy among services is high, when the cost of imprecise communication is low and when the types of problems faced by the services are similar, so that the optimal codes are close and the distortion required by the common code is small.<sup>3</sup> Second, codes interact in a

---

<sup>2</sup>The assumption that codes are the result of some optimization process (rather than pure historical accidents) can be defended in two ways. The case studies in Section 4 show that organizations do affect, in a deliberate way, the internal codes they use. At a more theoretical level, Rubinstein (2000) studies optimal languages and presents a model in which they arise as outcomes of evolutionary models. The literature review in Section discusses the connection to Rubinstein’s work.

<sup>3</sup>Trade-offs between coordination and specialization are also central in recent work by Hart and Moore (2005), who study how to allocate authority over the use of assets when agents with

meaningful way with the ‘vertical’ structure of the firm, and affect in particular the use of hierarchy. A hierarchical superior in our model functions as a translator, who enables services with different codes to cooperate. Thus hierarchies provide an alternative method for coordinating two services – common codes (associated with horizontal communication) and hierarchies are substitutes.<sup>4</sup>

Our analysis sheds some light on the impact of changes in communication costs on organization. As communications costs decrease, we expect to observe horizontal integration, as the cost of common codes decrease and capturing synergy becomes relatively more attractive. We also expect to observe vertical disintegration, as ‘translators’ are less necessary and units can communicate directly with each other. Our theory thus generates the implication that advances in information technology should result in an increase in common codes, an increase in peer-to-peer communication, a reduction in the number of layers of management within existing hierarchies or “delaying” and a broadening of firm scope. Thus our analysis provides a rationale for recent empirical findings on IT and organizational change, which we examine in Section 4, that suggest an increase in decentralization and a reduction in the number of layers.

More importantly, the theory reconciles two recent trends observed in organizations—towards information centralization and towards hierarchical decentralization, the so-called “empowerment”. As we argue in Section 4, one would expect that if the information is available centrally, then decisions should be made centrally. And yet, as we document in our case studies, it appears that the opposite is the case. Our analysis suggests that what is crucial is the homogeneization and standardization of categories across the organization, which facilitates horizontal communication between agents and allows for the substitution of hierarchical communication by horizontal communication mediated by a common code. We present two specific examples of these changes: the Microsoft corporation, where common human resource and finance categories were adopted throughout the organization, leading, according to the account of Robert J. Herbold, its Chief Operating Offi-

---

several assets (coordinators) can have ideas involving the common use of several of these assets, and when agents are motivated by their own interest rather than the organizations. The key comparative static prediction in their work concerns how changes in the importance of synergies affect integration and hierarchy. Our analysis generates similar comparative statics on integration and on hierarchy (where hierarchy is unrelated to authority) but also, unlike theirs, allows us to discuss the impact of information costs on horizontal and vertical structure since communication among agents is possible in our work but not in theirs. See also related work by Hart and Holmstrom (2002), where integration leads to more synergies, but at the cost of ignoring the private benefits of individual managers too much.

<sup>4</sup>We will not study the other obvious important issue in this framework, which is the optimal richness of codes and the trade-offs between investment in learning a new code and spending more resources on production activities. Garicano (2000) provides an example of the study of a similar tradeoff. The agents can learn how to process more tasks, but this has a cost.

cer for Microsoft from 1994 to 2001, to an increase in decentralization; and the development of the B-2 project, where the traditional between-team vertical communication (mediated by higher up ‘translators’) was substituted by horizontal, peer-to-peer, communication thanks to a common set of categories across teams.

Section 5 discusses links with this and other previous literature, and directions for future research. All proofs are in Appendix A. Appendix B extends some results presented in the body of the paper.

## 2 A Simple Theory of Language

In this section, we lay down a simple theory of the choice of a technical language or code, beginning with the case of two agents who need to communicate with each other, in subsections 2.1 to 2.3, and turning to the case of an agent who needs to communicate with several other agents in subsection 2.4. Finally, in subsection 2.5 we apply our results to a simplified environment, which will be used in the rest of the paper.

### 2.1 Model

A “salesman” confronts a problem  $x$ , drawn with probability  $f_x$  from a finite set  $X$ .<sup>5</sup> An expert or “engineer” can solve the problem, but this requires diagnosis (which will be defined shortly). Solving the problem is trivial for the engineer, once the problem is diagnosed. Salesmen can classify problems, but not perfectly, since they are boundedly rational. A code is a shared technical language that allows the salesman to transmit coarsely the information he has obtained about the problem. Formally, a code  $\mathcal{C}$  is a partition  $\{W_1, W_2, \dots, W_K\}$  of the set  $X$ . By uttering the word  $k$ , the salesman lets the engineer know that the problem  $x$  belongs to the subset  $W_k$ . We speak about the *breadth* of word  $k$ , the number of events  $n_k \equiv \#W_k$  that it contains, and about its frequency,  $p_k \equiv \sum_{x \in W_k} f_x$ .

The bounded rationality of the agents is represented by the maximum number of words,  $K$ , that they can learn.

Having received word  $k$  from the salesman, solving the problem requires that the engineer identify the precise problem  $x$  of the client. This *diagnosis* stage takes time and/or energy, and such cost is larger the broader the word. In particular, the diagnosis cost is an increasing function  $d$  of the breadth  $n_k$  of  $k$ . The

---

<sup>5</sup>Nothing in our analysis requires a finite set, but the exposition is simplified by this assumption.

expected diagnosis cost of code  $\mathcal{C}$  is therefore

$$D(\mathcal{C}) = \sum_{k=1}^K p_k d(n_k). \quad (1)$$

If, as we shall assume until further notice, the value of serving a client is high enough that the organization would never want to exclude solving specific problems with certain values of  $x$  just for the sake of saving on diagnosis costs, the profit maximizing code is the code that minimizes the expected diagnosis cost.

The following simple example makes our definition of diagnosis cost more concrete. The two agents are medical doctors: the salesman is a general practitioner and the engineer is a specialist. The problem consists in diagnosing a patient who presents a number of symptoms. The general practitioner sees the patient first, may run a battery of tests, and then prepares a referral for the specialist. The referral describes, using the specialized ‘code’ the knowledge that the generalist has gained about the patient’s problem. To the extent that the code is imprecise, due to the agents’ bounded rationality, the specialist will spend more time and effort diagnosing the patient’s problem. Such cost is a diagnosis cost (literally in this case – but it also corresponds to our definition). The more imprecise the referral, the higher the diagnosis cost.

Throughout this section, we take the frequency of events that agents deal with as exogenous. Of course, the frequency of events is partly a consequence of organizational design decisions. In Section 3 below we endogenize such organizational design decision and thus the actual distributions of events agents’ confront.

## 2.2 Optimal codes

In this subsection, we derive some properties of the optimal code that two agents would use.

**Proposition 1.** *In an optimal code, broader words describe less frequent events: if  $n_k > n_{k'}$ , then  $f_x \leq f_{x'}$  for any  $x \in W_k$  and  $x' \in W_{k'}$ .*

Proposition 1 implies that events of similar frequencies are grouped together in an optimal code: if  $f_x < f_{x'} < f_{x''}$  and  $x$  and  $x''$  belong to the same word, then  $x'$  also belongs to that word. Intuitively, agents use more precise words for those events they confront more often– like in our example of Hollywood types having categories determining whether certain actors can or cannot open a movie, or economists having categories for specific types of strategies.

Proposition 1 relates word breadth to event frequency. The next result, Proposition 2, relates word length to word frequency; it requires the additional assumption that the function  $d$  is (weakly) convex.

**Proposition 2.** *Assume*

$$d(n + 1) - d(n) \geq d(n' + 1) - d(n') \quad (2)$$

*whenever  $n \geq n' \geq 1$ . Unless integer constraints make it impossible, in an optimal code broader words are used less frequently: if  $n_k - n_{k'} \geq 2$ , then  $p_{k'} \geq p_k$ .*

Intuitively, to see that broader words are used less frequently: if  $n_k \geq n_{k'}$ , then  $p_{k'} \geq p_k$ , start from an optimal code and suppose we transfer an infinitesimal event  $x^*$  from a broad word to a narrower word. After transferring it, the event  $x^*$  is captured by a narrower word, and this is a certain benefit. Moreover, the broad word is now less broad and the narrow word is less narrow. If the broad word were used more often, this would also yield a benefit; but the original code was optimal, thus it cannot be a benefit, and the broad word must be used less frequently.

We have assumed that events could be allocated between words arbitrarily. In some instances, however, the “meaning” of events imposes constraints on the languages which can be constructed. For example, if we are partitioning the color spectrum into discrete color words, most words will group contiguous points of the spectrum. In Appendix B we extend Propositions 1 and 2 to environments where events have a natural ordering: we show that for two contiguous words, the broader word is used less often and describes events with a lower average frequency.

### 2.3 Environment Complexity

Firms face environments with varying degrees of complexity, and this affects the value of the codes that they use. For example, a bakery has always the same type of problem— a client comes in and wants to order one of several possible cakes; while an auto repair shop has to deal with a wide range of problems having to do with all possible malfunctions. The bakery environment is low complexity, while the auto repair shop is high complexity. In this subsection, we explore how the complexity of the environment affects the optimal diagnosis cost.

In our setting, the environment in which the firm operates is characterized by the distribution of events it faces. We consider two distributions of the same set events,  $f$  and  $\tilde{f}$ . The environment generated by  $\tilde{f}$  is more *complex* than the environment generated by  $f$  if, for any  $n$ , any word that contains the  $n$  events with the smallest frequency according to  $f$  has a greater probability according to  $f$  than to  $\tilde{f}$ . More precisely:<sup>6</sup>

---

<sup>6</sup>Note that the complexity of the environment defines an order on distributions. Of course, these definitions are very close in spirit to first order stochastic dominance.

**Definition 1.** *Density function  $\tilde{f}$  generates a less complex environment than density function  $f$  if  $f_W \geq \tilde{f}_W$  for any word  $W$  such that  $x \in W$  and  $x' \notin W$  implies  $f_x \leq f_{x'}$ . If for such a word  $f_W > \tilde{f}_W$ , then  $f$  generates a strictly less complex environment than  $\tilde{f}$ .*

With this definition, the communication cost is increasing in complexity:<sup>7</sup>

**Proposition 3.** *If  $\tilde{f}$  generates a less complex environment than  $f$ , the minimal diagnosis cost associated with  $\tilde{f}$  is (weakly) smaller than the minimal diagnosis cost associated with  $f$ .*

In a simple environment, there are a few extremely likely events and a large number of rare events. Communication costs are low, because the optimal code assigns likely events to narrow words, and narrow words are very probable. The worst-case scenario occurs when all events are equiprobable: words will divide the event space into equiprobable sets, and this will impose a high communication cost.

An immediate consequence of proposition 3 is that there exists an interesting interaction between the benefit of adding words and the complexity of the environment in which the firm operates.

**Proposition 4.** *Increasing the number of words from 1 to  $K > 1$  lowers communication costs more for less complex environments. On the other hand, moving from  $K$  words to a very large number of words (perfect communication) lowers communication more for more complex environments.*

Having a code with more words is always useful, but its relative benefit depends on how complex the environment is and how rich the language already is. Starting from the coarsest language, adding words is most beneficial in a simple environment. The savings in terms of diagnosis cost are high because few words can describe precisely a large share of events. If instead the code is already quite rich, the additional cost reduction is greater for complex environments, simply because the diagnosis cost is still considerable.

Our analysis points out to a general non-monotonicity in the size of the code a firm adopts, in relation to environment complexity. If words are expensive, we may observe a basic code for simple environments but no code at all for complex ones. If the cost of words goes down, the code will stay quite basic in the former case, but it can jump from non-existing to very rich in the latter case.

---

<sup>7</sup>The assumption that the diagnosis cost is convex is not needed.

## 2.4 Shared codes and dialects

The organizational analysis of section 3 will rely heavily on the insight that communications between organizational units require the use of a common code. In this subsection, we provide some background for this analysis by studying the choice of code by an engineer who needs to communicate with two<sup>8</sup> salesmen,  $A$  and  $B$ , who face the same set of events  $X$  but have different distributions  $f_x^A$  and  $f_x^B$ . The stark model which we are using implies that there are no “dialects”. At the end of this subsection, we explain how it could be enriched in order to allow for their existence. We also argue that this should not affect our main insights.

There are three agents, the engineer and salesmen  $A$  and  $B$ , who face the same set of events  $X$  but have different distributions  $f_x^A$  and  $f_x^B$ . Per period, salesman  $A$  receives requests from  $m_A$  clients, and salesman  $B$  requests from  $m_B$  clients.

Each of the three agents can learn at most  $K$  words. Agent  $A$  uses code  $\mathcal{C}_A$ , agent  $B$  uses code  $\mathcal{C}_B$ , and the engineer, who must understand both agents, knows code  $\mathcal{C}_A \cup \mathcal{C}_B$  (of course, he only uses the relevant part of this code when communicating with a salesman).

For instance, with  $X = \{1, 2, 3, 4, 5, 6\}$ , we could have

$$\mathcal{C}_A = \{\{1, 4\}, \{2, 5\}, \{3, 6\}\}, \quad (3)$$

$$\mathcal{C}_B = \{\{1, 2, 3\}, \{4, 5, 6\}\}. \quad (4)$$

Then, the engineer must know five words, while  $A$  knows 3 and  $B$  knows 2.

Proposition 5 shows that the same code, which saturates the rationality constraints of all the agents, will be used in communicating with both salesmen.

**Proposition 5.** *The optimal codes contain  $K$  words and satisfy  $\mathcal{C}_A = \mathcal{C}_B$ .*

Two examples can illustrate the proof of proposition 5. First, with  $K = 5$ , consider the codes of equations (3) and (4). The narrowest noncommon words<sup>9</sup> are  $\{1, 4\}$ ,  $\{2, 5\}$ , and  $\{3, 6\}$ ; let us introduce  $\{1, 4\}$  into  $\mathcal{C}_B$ . Then,  $\tilde{\mathcal{C}}_B = \{\{1, 4\}, \{2, 3\}, \{5, 6\}\}$ ; every event is now represented by a shorter word, and diagnosis cost must go down while the engineer must still learn five words,  $\{1, 4\}$ ,  $\{2, 3\}$ ,  $\{5, 6\}$ ,  $\{2, 5\}$  and  $\{3, 6\}$ . Notice that  $\mathcal{C}_A$  and  $\tilde{\mathcal{C}}_B$  are still not efficient: if we add  $\{2, 5\}$  to  $\tilde{\mathcal{C}}_B$ , the engineer must still learn five words ( $\{1, 4\}$ ,  $\{2, 5\}$ ,  $\{3\}$ ,  $\{6\}$  and  $\{3, 6\}$ ), and the codes are more efficient.

<sup>8</sup>It should be clear that the assumption that there are two agents is made only for ease of exposition and is totally unnecessary for the results.

<sup>9</sup>We are taking some liberty with our terminology. Strictly speaking, we have defined a word to be the *name* of a set of events. In this discussion, a word is the set of events itself. This should create no confusion, and lighten considerably the exposition.

A more complicated example starts from

$$\begin{aligned} \mathcal{C}_A &= \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9, 10\}, \{11, 12, 13, 14, 15, 16\}\}, \\ \mathcal{C}_B &= \{\{1, 4, 7, 11\}, \{2, 5, 8, 12\}, \{3, 6, 9, 13\}, \{10, 14, 15, 16\}\}. \end{aligned}$$

If we take  $\{1, 2, 3\}$  as the narrowest noncommon word, the new code for  $B$  is

$$\tilde{\mathcal{C}}_B = \{\{1, 2, 3\}, \{4, 7, 11\}, \{5, 8, 12\}, \{6, 9, 13\}, \{10, 14, 15, 16\}\}.$$

All events but 10, 14, 15 and 16 are now represented by shorter words.

We have shown that the engineer will use the same code to speak to both salesmen. Which code will be chosen? It is easy to see it will be the code which would have been chosen had the engineer faced only one salesman with a distribution of events equal to the expected distribution of events for the two salesmen. This is formalized in the next proposition.

**Corollary 1.** *Propositions 1 and 2 apply as stated to the common code if one defines*

$$f_x = \frac{m_A f_x^A + m_B f_x^B}{m_A + m_B}.$$

Of course, in reality, we would expect the engineer (and more generally hierarchical superiors) to know more words than salesmen, for two reasons. First, if bounded rationality imposed a cost on the acquisition of language rather than an absolute limit on the number of words that can be learned, it would be optimal for the engineer to learn more words than the salesmen. Second, if some agents have different abilities, it would be optimal to choose agent who is able to learn the most words as the engineer. Presumably, it would be optimal for the salesmen to share some words, while using specific words to communicate to the engineer events that they encounter much more often than the other salesman.

Garicano (2000) and Garicano and Rossi-Hansberg (2003) build theories of hierarchies where agents who have the ability to complete more tasks are chosen as “supervisors”; it may be possible to build similar theories, where the supervisors are able to learn more words. We take a small step in that direction in 3.3, where we assume that the firm can hire, at an additional cost, an agent with a larger  $K$ .

While our result on the suboptimality of dialects is clearly quite special, we believe that the intuition behind it is general and of great economic significance. Two organizations can either develop a common language or not. If they do, they pay a certain cost to have all their members share the same code but then they reap high rewards from smooth communication. If they keep separate languages, they enjoy the advantage of being able to tailor each of them to the their specific

needs. A half-way solution based on a partially shared language can be very costly in terms of the need to learn new words (or abandon words with precise meanings) and it can offer very limited rewards in terms of synergies, because communication between organization is still difficult.

The shortcomings of dialects can be illustrated with an example close to home. Scientists who engage in joint projects with other disciplines are often painfully aware of the difficulty of short-term interdisciplinary work: an inordinate amount of time is spent learning bits and pieces of the other side’s language, and still communication is slow and full of misunderstandings. The interdisciplinary project pays off only in the long term – when the two sides have managed to develop a joint language that allows them to access each other’s knowledge in a seamless matter. Interdisciplinary work should either be at a very basic level, where communication is infrequent and occurs through common language (an economist can read a popular physics book) or it should involve a serious long-term institutional effort to create a common code (as in the case of the disciplines that are taught in business schools, or more recently in the case of long-term collaborative efforts between psychologists and economists).

## 2.5 Two Word Codes

To facilitate the organizational analysis that follows, we simplify the framework which we have used up to this point: we consider two word codes, a linear diagnosis cost, and assume a continuum of events.

Each salesman deals with consumers  $x \in [0, 1]$  with cumulative distribution functions

$$F(x; b) = (1 - b)x + bx^2,$$

and density

$$f(x; b) = (1 - b) + 2bx,$$

where  $b \in [-1, 1]$  measures the inequality of the distribution of events.

The diagnosis cost is linear: if the engineer knows that the client’s characteristics fall in an interval of length  $a$ , his diagnosis cost is simply  $\lambda a$ , where  $\lambda$  is a positive parameter.

With an engineer and a single salesman, we can adapt proposition 1 to show that they will use a “right word” and a “left word”, separated by  $\hat{x}(b)$  which minimizes

$$s(x) = F(x; b)x + (1 - F(x; b))(1 - x), \quad (5)$$

which implies,<sup>10</sup>

$$\hat{x}(b) = \operatorname{argmin}_x s(x) = \frac{1}{6b} \left( 3b - 2 + \sqrt{(3b^2 + 4)} \right). \quad (6)$$

The corresponding expected diagnosis cost is

$$\lambda D^*(b) = \lambda \times \frac{8 + 36b^2 - (4 + 3b^2)^{\frac{3}{2}}}{54b^2}. \quad (7)$$

The variations of  $D^*$  and  $\hat{x}$  as a function of  $b$  are represented on figure ???. The diagnostic cost is convex in  $b$ , and is maximal for  $b = 0$ . The cutoff event  $\hat{x}$  increases close to linearly in  $b$ .

### 3 Language and Organization

In the previous section, we studied the optimal code for exogenously given organizations. In this section, we optimize jointly on organizational structure and code: who should communicate with whom? what code should they use?

We develop a simple model with two services, each composed of one salesman and one engineer. We shall study communication and coordination among the two services, focusing on three possible organizational forms: *separation*, where the two services use different codes; *integration* where the two services share the same code; *translation*, where there exists a hierarchical structure supplying an interface between the services. We will identify the environments in which each of these organizational forms is optimal.

#### 3.1 Synergies

We assume that there is some form of synergy between the two services, which gives rise to potential benefits of coordination. Specifically, we assume that if the two services do not communicate with each other, they can each deal with a quantity  $q^{NC}(p)$  of clients, while if they communicate they can serve  $q^C(p)$  clients. The parameter  $p$  measures the intensity of the potential synergy between the two services.

Two assumptions are natural. First, for any level of synergy  $p$ , more clients are served if there is communication:  $q^C(p) > q^{NC}(p)$ . Second, there is complementarity between the level of synergy and the benefit of communicating: an

---

<sup>10</sup>The function  $s$  is not convex in  $x$ . However, the quadratic function  $s'(x)$  is convex with  $s'(0) < 0$ . Hence, there exists a unique  $\hat{x}(b) \in (0, 1)$  such that  $s'(\hat{x}(b)) = 0$ ,  $s(x)$  is decreasing on  $[0, \hat{x}(b)]$  and increasing on  $[\hat{x}(b), 1]$ . Profits are single peaked, with a maximum at  $\hat{x}(b)$ .

increase in the intensity of synergy  $p$  makes communicating services relatively more productive than non-communicating services. Formally, if  $p'' > p'$ , then

$$\frac{q^C(p'')}{q^{NC}(p'')} > \frac{q^C(p')}{q^{NC}(p')}. \quad (8)$$

This formulation is agnostic as to whether an increase in synergy is due to a generally positive event ( $q^C(p'') > q^C(p')$  and  $q^{NC}(p'') > q^{NC}(p')$ ), a generally negative event ( $q^C(p'') < q^C(p')$  and  $q^{NC}(p'') < q^{NC}(p')$ ), or one that is positive only for communicating services.

**Example 1.** One simple example of synergy that fits our framework is the management of excessive capacity. Services can help each other deal with overflows of clients. Each service has limited capacity: in each period, it can accommodate only one client. On the other hand, the number of clients who knock at its door is random, it can be zero, one or two. With positive probability, there is excessive load in one service and excessive capacity in the other. When this happens, the firm benefits from diverting some business from the overburdened service to the other.

Formally, in the example each engineer has the ability to attend to the needs of at most one client. The client arrival process is as follows (see Figure ??):

$$y = \begin{cases} 0 & \text{with probability } p, \\ 1 & \text{with probability } (1 - 2p), \\ 2 & \text{with probability } p, \end{cases}$$

where  $p$  belongs to the interval  $[0, 1/2]$ . This arrival process captures the effect of the variability in the expected number of clients of each type. If  $p$  is low, then each salesman is likely to find one client per period of each type. When  $p$  is high, although on average still 1 client is arriving, it is quite likely that either none or 2 will arrive. Thus  $p$  measures the importance of the synergy between the two services: a high  $p$  means that the services are likely to need to share clients, while a low  $p$  means that each service is likely to have its capacity fully utilized. This form of synergy between services satisfies our assumptions. Non-communicating services are able to deal with an expected number of clients

$$q^{NC}(p) = 2((1 - 2p) + p) = 2(1 - p),$$

while communicating services can also cover the case where one service receives zero problems and the other one receives two, which happens with probability  $2p^2$ :

$$q^C(p) = 2(1 - p + p^2).$$

It is easy to check that this formulation satisfies the two requirements on  $q^{NC}$  and  $q^C$  that we imposed above.

The two salesmen face different distributions of consumers, which, for simplicity, are assumed to be symmetric to each other. Salesman  $A$  face consumers whose characteristics are drawn from the distribution  $F(x; b)$ , while salesman  $B$  faces distribution  $F(x; 1 - b)$ . The parameter  $b \in [0, 1]$  measures the inequality of the probability of different events for each salesman. Note that the distribution of characteristics of clients over all the clients who approach the firm is  $F(x; 0)$ .

The profit of the firm when it solves a client's problem is 1. The per-client diagnosis costs is  $\lambda$ : if the engineer knows that the client's characteristics fall in an interval of size  $s$ , his diagnosis cost is  $s\lambda$ . We assume that the diagnosis cost is sufficiently high to ensure that information must transit through a salesman before being sent to an engineer ( $\lambda > 1$ ) but not so high that profit risks becoming negative ( $\lambda < 2$ ).

### 3.2 Integration or separation?

In this section, we study the choice between segregation of services, where the salesman from service  $i$  only communicates with the engineers of his service, and integration where a salesman can communicate with either engineer.

In an *integrated organization*, a salesman must be able to communicate to the engineer from the other service the needs of his customer: indeed, as seen above, because  $\lambda > 1$  sending the problem to the engineer without explanation is not profitable. The proof of proposition 5 can easily be adapted to show that the two services must use the same code, and the proof of corollary 1 can be adapted to prove that the optimal common language is  $\hat{x}(0) = \frac{1}{2}$ , the language which would be optimal with one engineer receiving messages from one salesman for which the density of characteristics is the average density of the two services,  $F(x; 0)$ . Total profits are

$$\Pi^I = q^C(p)(1 - \lambda D^*(0)), \quad (9)$$

where  $1 - \lambda D^*(0) = 1 - \frac{1}{2}\lambda$  is the expected total profit from serving one customer.

By (7), the maximum total profit in a *separated organization* is

$$\Pi^S = q^{NC}(p)(1 - \lambda D^*(b)). \quad (10)$$

It is easy to check that the profits of a separated organization are positive.

We can now study the comparative statics of the choice between integration and separation:

**Proposition 6.** *An integrated form becomes relatively more profitable compared to the segregated form when a) the diagnosis cost  $\lambda$  decreases, b) the synergy parameters  $p$  increases, or c) the heterogeneity of the two client distributions  $b$  decreases.*

All three parts of Proposition 6 are intuitive. If the diagnosis cost parameter increases, it becomes more costly to rely on the imprecise language associated with an integrated form, and the organization may prefer to forgo the benefits of synergy in order to save on diagnosis. A higher level of synergy between services will clearly make it more profitable to have the two services communicate. Finally, when the problems that the two services face become more similar, there is less difference between the optimal common code and the two optimal specialized codes. The additional diagnosis cost associated with an integrated structure is smaller and having a joint code becomes more profitable.

### 3.3 Hierarchy

Empirical studies have noted that organizations often use specialized “translators” to facilitate communication between the inside and outside of the organizations. For instance, Tushman (1977), in a study of a large R&D facility of a Midwestern Corporation, finds that a “small set of individuals able to translate between coding schemes” function to link the innovating system with various sources of external information.<sup>11</sup> We consider here such an organization, where the two services use different codes but exploit the synergy by employing a fifth agent who provides translation. When inter-service communication is needed, the translator steps in. If salesman  $A$  has a problem for engineer  $B$ , he communicates to the translator the type of the problem in the code used in service  $A$ . The translator will search

---

<sup>11</sup>Allen and Cohen (1969) provide the first systematic evidence of the role of translators in facilitating communication between organizations. In a study of inter and intra-lab communication in R&D labs, they find that the average agent is “connected by a specialized intermediary to information sources outside its lab”. A relatively small number of specialized “gatekeepers” (in their language) undertake the majority of the outside communication of the lab and obtain most of the outside information that is relevant to the lab’s technical work. Consistently with our interpretation of these “translators” as managers, Allen and Cohen (1969) find that both of the lab research directors and two thirds of the first line supervisors were among these sources of key external information. Allen (1970) obtain similar findings in a large aerospace firm. For a discussion of this literature, see e.g. Ancona and Caldwell (1992); and Kogut and Zender (1992).

for  $x$ , and then he will transmit the information to engineer  $B$  in the code used in service  $B$ .

Hiring a translator requires incurring a fixed cost  $\mu$ , but since the translator is specialized in language, we assume that his diagnosis cost is lower than that of the engineers. In particular, we assume here that this cost is zero. The qualitative results of the analysis are valid if it is strictly positive, as long as it is lower than the cost of diagnosis for engineers.

Engineer  $A$  now receives problems from salesman  $B$  via the translator and from salesman  $A$  directly. The frequency with which he receives a problem from the other service's engineer is an increasing function  $\phi(p)$  of the level of synergy.

**Example 1. (cont).** In the excessive capacity example, the function  $\phi$  is

$$\phi(p) = \frac{p^2}{1 - p + p^2},$$

which is increasing for  $p \in (0, \frac{1}{2})$  and takes values between 0 and  $\frac{1}{3}$ .

The optimal code with translation differs from the optimal codes of both the integrated form and the separated form. The optimal code for engineer  $A$  is based on problems that come with probability  $(1 - \phi(p))$  from a distribution with parameter  $b$  and with probability  $\phi(p)$  from a distribution with parameter  $-b$ . In our set-up this corresponds to a distribution with cumulative distribution function  $F(x; (1 - 2\phi(p))b)$ . The optimal code for form  $H$  has diagnosis cost  $D^*((1 - 2\phi(p))b)$ , which is greater than the diagnosis cost for the separated form (because the environment faced by services is more complex) and smaller than the cost for the integrated form (because the constraint that the services use the same code has been relaxed).

The following proposition describes the variation of the optimal organization as a function of  $\lambda$ .

**Proposition 7.** *For any  $b$ , if  $p$  and  $\mu$  are high enough, there exist  $1 \leq \lambda_{\min} < \lambda_{\max} \leq 2$  such that the unique optimal organization is*

$$\begin{array}{ll} \text{integrated} & \text{if } \lambda < \lambda_{\min} \\ \text{hierarchical} & \text{if } \lambda \in (\lambda_{\min}, \lambda_{\max}) \\ \text{separated} & \text{if } \lambda > \lambda_{\max} \end{array}$$

**Example 1. (cont).** Figure ?? shows how the optimal organizational form varies as a function of the synergy parameter  $p$  and the diagnosis cost parameter  $\lambda$  in the excessive capacity example. Consider first the choice between translation (*i.e.*, hierarchy) and separation. Translation incurs a fixed cost  $\mu$  and increased diagnosis costs,

but makes inter-service communication possible and thus allows the services to profit from the existing synergies. If the diagnosis cost  $\lambda$  is low, the extra cost due to translation is low and the net benefit is likely to be high. Thus, translation is more likely to be preferred to separation when  $\lambda$  is low.

Translation allows the two services to keep efficient service-specific codes. It is likely to be preferred to integration when well adjusted codes are more important, that is when  $\lambda$  is large. As stated by the proposition, if the fixed cost  $\mu$  of hiring a translator is low enough, for  $\lambda$  large enough the hierarchical structure is preferred to integration. However, if the diagnosis cost parameter is too large, then the communication is simply not worth it and the optimal structure is the separated form.

## 4 IT Revolution, Technical Languages, and Organizational Change

In this section, we use our model to interpret systematic evidence that information in organizations is becoming increasingly centralized but decision-making becomes more decentralized. Such changes are paradoxical in either a pure agency or information perspective, since increasing information centralization should lead decision rights to be more centralized in both of these perspectives. The whole point of tolerating agency problems is to rely on agents local knowledge to make decisions; if information is centralized, then agency conflicts can be reduced at low cost by centralizing decision making. Similarly, decentralized information processing makes sense to gather information from agents; if information is centrally located, agents role should diminish, rather than increase. Our theory, through the substitution of hierarchy by common codes, makes the opposite prediction in a natural way. We then breathe life into the link between theory and evidence by looking at two important cases: the re-organization of Microsoft and the adoption of a common code for the construction of the B-2 Bomber.<sup>12</sup>

**Theory Implications: Diagnosis Costs Comparative Statics.** In the past two decades, rapid advances in information technology have caused a dramatic drop in a range of information costs, including the cost of communicating information, of accessing information contained in data bases and of processing information. For the purpose of our analysis a key feature of these changes is that by

---

<sup>12</sup>A separate empirical literature documents the impact of different natural (rather than technical) languages on trade. It finds that a common language has an important positive impact on between country trade. See Melitz (2005) for evidence and references to previous work.

reducing the cost of searching information, these IT advances reduce the cost of receiving an imprecise message from a third party. The reason is that the receiver can more cheaply conduct the follow up diagnosis of the problem. If, for example, a generalist describes the illness of a patient in general terms such as: “I believe the patient has a pulmonary illness that was discussed in a recent issue of JAMA” the specialist can actually figure out what he must mean specifically with a quick key word search of the relevant database (*Medline*).

Our theory predicts that these changes should imply an increase in integration in the form of links across and within firms, through the use of hierarchies and common codes; moreover, within already integrated units, decreases in diagnosis costs reduce the translation role of hierarchy, by facilitating horizontal communication – the substitution of common codes for hierarchies and increase in peer-to-peer communication. As we have argued in the introduction, these predictions – horizontal integration and vertical disintegration– are specific to our model, since it is the first in the literature that allows for both horizontal (between similar agents) and vertical agents. Both of these broad predictions appear consistent with the evidence.\

**Systematic Evidence.** First, the reduction in information costs is correlated with increasing code commonality. Historically, the information generated by each business unit within a firm and by each function within each business unit has been coded and processed separately, according to the needs of that business unit or function; the different pieces of information were often defined in different ways and could not be easily aggregated.<sup>13</sup> As information costs have dropped, companies have sought ways to integrate this disperse information. In particular, this integration was obtained, between and within firms, through tools such as Enterprise Resource Planning (ERP) systems<sup>14</sup> and, earlier, Electronic Data Exchanges (EDI),<sup>15</sup>. This programs allow for the exchange of electronic data by standardizing the format of the data exchanged. Through these systems, firms

---

<sup>13</sup>For example the database company Oracle had 70 incompatible databases for its human-resources department. This made it impossible to answer simple queries, such as how many employees were working at any time at the company. “If anyone wanted to find out the exact number of Oracle employees, it would take weeks of searching— and by the time the answer was found, it would already be out of date.” (“Timely Technology,” *The Economist*, January 31, 2002.)

<sup>14</sup>See for instance the products offered by SAP <http://sap.com> or Baan <http://www.baan.com>.

<sup>15</sup>We refer to EDI systems broadly, to include other related approaches such as CPFRR (“Collaborative Planning, Forecasting and Replenishment”) which involves deeper and more extensive electronic information sharing and has been installed, for example, by Nabisco and used with Webman’s Food markets (“Enterprise System,” *Financial Times*, February 22, 1999); or web-based integrated value chains, such as the one introduced by Safeway in the UK (“You’ll Never Walk Alone,” *The Economist*, June 24, 1999).

have substituted flexible ways to code their data by more rigid but unified central databases.<sup>16</sup> These common information systems have resulted in increasing horizontal information links within and between firms, as our examples below show.<sup>17</sup>

Second, the reduction in information costs induces greater decentralization. Brynjolfsson and Hitt (2000) were the first to find evidence of this complementarity between IT and decentralization. Bresnahan, Brynjolfsson and Hitt (2002) find, using firm-level data, that greater use of information technology is associated with broader job responsibilities for line workers, and more decentralized decision-making. Caroli and Van Reenen (2001) also find, on entirely different data, evidence that the degree of decentralization of authority is complementary with the use of IT. Rajan and Wulf (2003), in a panel study of the hierarchical structure of firms, find that the span of control of the CEO is increasing over time, in particular, through the disappearance of the role of the COO. With more employees under his direct authority, the CEO can exert less control: decision making is more decentralized.

Thus, the evidence does suggest that the drop in information costs led to (1) increasing commonality of codes in organizations and (2) increasing decentralization at the expense of hierarchy. This is not sufficient to show that these changes are causally linked in the same way as our model describes, that is, that the introduction of a common code allows for the substitution of the ‘translation’ role of hierarchy by direct horizontal communication between business units that otherwise would be ‘speaking a different language.’ We use some case-study evidence to illuminate this connection.

**Microsoft.** Robert J. Herbold,<sup>18</sup> Chief Operating Officer of Microsoft at the time, explains that in 1994 Microsoft had a completely decentralized set of information systems (Herbold (2000)). Each business unit used a different mapping of data to categories: in the terminology of this paper, they all used different, specific, codes. For example, the financial managers of the different units had chosen their own categories in their financial reports, adapted to their own circumstances. In Herbold’s words:

“Some would develop financial information systems tailored to

---

<sup>16</sup>In the words of a ‘noted American e-commerce expert’ cited by *The Economist*, ERP systems have replaced “fragmented unit silos with more integrated, but nonetheless restrictive enterprise silos” (“Timely Technology,” *The Economist*, January 31st, 2002).

<sup>17</sup>Future work is needed on whether the horizontal scope of organizations has been increased by IT. We discuss this in the conclusions.

<sup>18</sup>We rely heavily on Herbold personal account, in his *Harvard Business Review* article. All the quotes below are from his account.

their particular needs. Others would analyze their financial performance in a way meant to reflect the environment of their country of operation. There was nothing seditious about this.”

An example of these needs is given by the German country manager:

“We put years into the development of our own information systems because those systems uniquely capture the nuances of the German Business. Those nuances are important.”<sup>19</sup>

Similarly, there was no way to have a coherent overall image of human resources throughout the firm, with eighteen HR-related databases all using different ways to categorize the data.

“When asked about head counts, managers answers usually were, to put it charitably, poetic.”

The tailoring of the information to the specific needs of the different business units compromised communications between them, as different measures needed to be reconciled.

Taking advantage of the drop in information costs, Microsoft introduced ‘common codes’ in these two areas; now, according to Herbold, all managers could easily make sense of the information produced by any business unit, and thus make quick decisions. Paradoxically, and as our model predicts, this centralizing move provided “benefits usually associated with decentralization” as managers had easy access to relevant information and could use it directly in their exchanges with one another.

**The B-2 Bomber.** The adoption of a common code for the design of the B-2 bomber by four independent firms provides further evidence on the relationship between technology, code adoption and decentralization. The development of the B-2 bomber began in 1981.<sup>20</sup> Advances in information technology made it possible for Northrop, Boeing, Vaught (a division of LTV) and General Electric, the

---

<sup>19</sup>Obviously, these complaints only show that the center thought the codes were inefficiently different while the country managers thought that the codes were just appropriately adapted to their different environments. On the other hand, the center presumably cares both about coordination between countries and the profits within each country, whereas the country managers care mostly about local conditions. There is therefore at least some presumption that the center’s objective function is better aligned with the interests of the firm as a whole.

<sup>20</sup>The account that follows draws heavily on a detailed cased study by Argyres (1999). For background information on the B-2 and links to other sources of information, see the site of the Federation of American Scientists: <http://www.fas.org/nuke/guide/usa/bomber/b-2.htm>.

four companies in charge of the design, to create a common set of categories and a central database to facilitate the design of the bomber. A key element of the construction of a common database was the ‘B-2 Product Definition System’, essentially a common code, a “technical ‘grammar’ by which engineers and others conveyed information to each other.”<sup>21</sup> The development of the B-2 was the “first major aerospace program to rely on a single engineering database to coordinate the activities of the major subcontractors on a large-scale design and development project” (Argyres, 1999:163). The use of this database had two consequences. First, designers from different companies could participate jointly in the design – the existence of a common code allowed integration of several teams where before there was none possible, an effect illuminated by section 3.2. Moreover, this integration reduced the hierarchical coordination, since among the main consequences of the creation of a relatively rigid, unifying codes was an increase in decentralized decision making: “the technical grammar defined by the B-2 systems established a social convention which limited the need for a single hierarchical authority.”(Argyres 1999: 173). This is consistent with our description of the substitution of hierarchies by codes in 3.3.

## 5 Related literature and conclusions

The inspiration for this paper comes from Arrow’s idea of coding, proposed in his celebrated *The limits of organization* (1974). After discussing the endogenous development of codes within organizations, he identifies the trade-off between general codes that allow for wide communication and specialized codes tailored to the needs of particular organizations. By formalizing the concept of coding, we have proven a number of novel theoretical results, concerning the structure of optimal language, the suboptimality of dialects, and the relationship between language complexity and environment complexity. We have also explored the relationship, hypothesized by Arrow, between the choice of organizational language and the choice of organizational structure. This has led to testable implications on how changes in processing costs and other technological characteristics determine the choice between an integrated structure, a hierarchical structure, and a separated structure. These findings throw some light on the impact of recent drops in IT on organization and information centralization.

Other work in various disciplines has dealt with codes and technical languages.

---

<sup>21</sup>“This grammar was established through a highly-developed and highly standardized data formation and modeling procedures of the system, which laid down well-defined rules for communicating complex information inherent in the part design” (Argyres, 1999: 171). These rules included tight definition of 14 part families and “agreed upon modeling rules for defining lines, arcs, surfaces etc.” (Argyres 1999:169).

First, an area outside of economics, information theory (Shannon, 1948) studies optimal codes. However, because the questions posed are very different, so is the analysis. In particular, information theory is concerned with issues such as: representing the messages with sequences of binary digits (bits) that are as short as possible, what is the mean length of a message?; what is the capacity of different types of channels? what is the error in decoding, and what does it depend on?; can one design a code that allows as close to perfect communication as possible even in a noisy channel? In general, the theory assumes that the sender must transmit all of the information, and chooses the code that minimizes transmission cost. In our setting, which is concerned with the organizational implications of agents' bounded rationality, the transmission cost is given, but the sender is prevented from transmitting all the information. The optimal code maximizes the value of information transmitted. Moreover, we also study code adoptions by agents who can independently consider their own individual costs and benefits in adopting a code – this is obviously also out of the realm of information theory.

Rubinstein (2000) studies the economics of language from a number of perspectives. The closest to our approach is a model of optimal language (sections 1.3 and 1.4) where a planner is given a certain set of objects and can choose a binary relation to be imposed on the set. A binary relation is a generalization about the set. If the set is composed of authors, the relation could be: “Every author quotes another if he is older than himself.” This generalization may have exceptions (like, “author  $a$  is younger than  $b$  but does not quote  $b$ ”). The planner's goal is to find the most accurate binary relation, i.e. the generalization with the lowest number of exceptions. Rubinstein proves that the most accurate binary relation must be asymmetric and complete and that it will approximately split the set in two. Clearly, our approach differs from Rubinstein's on many dimensions: we do not use binary relations, we minimize diagnosis cost rather than maximize accuracy, we allow for more than two words, etc. However, at a deeper level we share Rubinstein's methodological approach: we describe optimal languages.

Rubinstein devotes a whole chapter (Ch. 2) to the issue of why we should be interested in optimal languages. He presents a model of language selection, with a static version and an evolutionary one. In the static version, multiple languages may emerge as equilibria and some of them are inefficient. However, the evolutionary model only supports the efficient language. The issue of language optimality is clearly a very important and difficult one. In this paper, we have chosen to look at efficient organizational languages mostly for a methodological reason: it allows us to describe the optimal choices organizations would make if they could. In practice, it would be important to know to what extent organizations are able to decide the code that their employees use.

A recent experimental literature, following Weber and Camerer (2003) aims to understand the way individuals create codes and the constraints such codes pose

on organizational success. In these experiments, individuals must create words to communicate quickly which picture they are looking at. Through repetition, they develop conventions that allow them to improve communication. The paradigm is different from ours, in that individuals can describe each event perfectly, but they aim to describe it fast; our approach is complementary, since we study individuals who have limited words, which are by necessity always imprecise descriptions, their objective is to maximize precision. This experiments allow Munyan and Camerer, (2005) to study mergers and find that the merged groups eventually reach the performance of the non-merged groups. Under our paradigm, that is when bounded rationality limits the size of the code, mergers must result, as our analysis in Section 3 showed, in permanent drops in communication performance. Further experimental work would be required to test this type of prediction.

Other antecedents of our work include Crémer (1993), who presents a bounded rationality analysis of corporate culture. He argues that ‘corporate culture is the stock of knowledge shared by members of the corporation, but not by the general population from which they are drawn’, and suggests that this knowledge stock is formed by three pieces: a shared knowledge of facts, a common code, and a shared knowledge of rules of behavior. He then goes on to study, within a team theoretic framework, the benefits of shared knowledge.<sup>22</sup> In the same context, Prat (2002) explores the connection between the optimal extent of information homogeneity within a team and the kind of complementarities that exist among different team members. However, neither of these two papers examine the design of codes or their consequences for the design of organizations. Battigalli and Maggi (2002) construct a sophisticated model of language, which they then use to develop a theory of contract incompleteness. Their language is a code with the purpose of legal verification which is built by combining primitive sentences and logical connectives (AND, OR, NOT, etc...). A contract uses the available language to partition the set of events and associate it to the parties’ obligations. Like Battigalli and Maggi’s we take into explicit account the cost of using language to partition the set of events. However, our focus on organizations is radically different from their focus on contracts.<sup>23</sup> Like us, Wernerfelt (2003) considers codes that minimize communication costs within an organization, but with a different focus. In his model, agents have common interests but decisions are decentralized, and he studies the existence of symmetric or asymmetric equilibria with one or multiple codes. His analysis does illuminate the existence of equilibria with different codes

---

<sup>22</sup>Chowdhry and Garmaise (2004) build on this work by studying organizational capital and corporate culture in a dynamic model of learning in the firm and analyze the consequences of such capital for financial magnitudes.

<sup>23</sup>Similarly in their focus on contracting and not in organization, Chatterji and Filipovich (2004) describe a language as a partition of the set of possible histories of the game generated by contracts and study the effect of language ambiguity on judicial interpretation.

when a common codes would be optimal, but does not extend to organizational implications or to the strategic aspects of the choice of codes. Finally, Dewatripont and Tirole (2003)'s study of communication emphasizes the strategic interactions between the communication efforts made by different agents, not the language they use.

Building on Marschak and Radner's (1972) *team theory* a number of authors have studied how the limits on communications imposed by affect organizational structure. Crémer (1980) studies the optimal allocation of tasks into divisions, whereas other authors have been more interested in developing a theory of hierarchies. Radner (1993) and others (see Van Zandt, 1999 for a survey) stress the limited computation capacity of agents. Closer to our work is Bolton and Dewatripont (1994), who consider a more general communication cost structure; this leads to a theory which builds on the trade-off between communication costs and returns to specialization. In Garicano (2000), the bounded rationality of agents prevents them from learning the solution to all the problems potentially faced by the organization, but they can request help from other agents when they do not know how to solve one of them. He shows that the firm will organize itself in a knowledge hierarchy, in which agents closer to the production floor deal with the most common problems while higher rank agents deal with less frequent problems.

Thus, to the best of our knowledge, none of the previous literature studies the relationship between the organizational code and the organizational choices of the firm. Specifically, our analysis illuminates at least three sets of issues absent from the literature. First, and at an immediately applied level, it provides an explicit link between centralized communication and code commonality. In particular, we have shown that codes and hierarchy are likely to be substitutes; that is, for a given firm scope, increasing code commonality will reduce the reliance on hierarchies. Second, more generally, we have provided a simple way to analyze and formalize an elusive idea, the idea of a code, and shown how this formalization could be used to study organization issues. Third, we have characterized some of the main strategic issues that are likely to bias code adoption.

Also, understanding the loss in precision imposed by the need for common codes offers a fresh, bounded-rationality based, perspective on organizational boundaries. Modern theories of the firm scope have emphasized the role played by different allocations of asset ownership in providing incentives to agents (e.g. Hart and Moore 1986). Our theory suggest that, if allowing agents to efficiently communicate with one another requires establishing a common code, then agents' bounded rationality limits the scope of the firm.<sup>24</sup> By including broader, more

---

<sup>24</sup>In fact, the original Coase (1937) paper on the firm's boundaries emphasized the importance of organization costs in setting a limit on firm size, and argued that a reason these organization

diverse business inside a common code, communication within each service deteriorates. Thus, the organization should group related services or products and segregate unrelated ones, since the latter cause larger decreases in the efficiency of communication with smaller gains in synergies.<sup>25</sup>

Our analysis suggests several interesting avenues for new research. First, our model yields testable hypotheses, and the availability of large databases of business texts and their ease of access may allow for a study of the commonality of the language used across different services of different firms or across different firms in an industry. Beyond testing the relation between integration of codes and characteristics of the environment, such research would allow for a direct test our hypothesis on the ‘centralized information, decentralized decision-making’ – the substitution of codes for hierarchies. In particular, one should observe more ‘de-layering’ (less hierarchy) and more horizontal communication as codes become more common.

Second, on a theoretical level, our analysis can be used to provide some structure to the concept of ‘hard’ and ‘soft’ information, which is increasingly used in the contracting literature (see Aghion and Tirole, 1997; Stein, 2002). These concepts can be given a precise meaning in our model. A word within a code is hard information; it can easily be passed far down the chain of command or in space. The exact meaning of the word, that is clarifying which of the possible events within it are referred to, is soft information – that is, it is not very costly to eventually figure out this meaning in one to one communication, as one can continue talking until the other party understands, but is very costly to do so when communication is long distance or along a chain of command.

Third, also on a theoretical level, it would be interesting to explore code adoption in a dynamic setting. We conjecture that there exists a U-shaped relationship between the persistence of the environment in which the organization operates and the persistence of the code that the organization uses. Codes are stable over time if the environment is either very immobile (a specialized code needs not be modified) or it is highly unpredictable (a constant non-specialized code is the best solution).

A fourth interesting extension would study how the analysis changes when individuals choice codes independently and act strategically. A previous version, available from the authors, studies such choices. Our analysis identifies a first

---

costs exist is the limiteds on the manager’s resources. Building on that insight, Oliver Williamson (1967) starts a tradition, which we discuss in the conclusions, of modeling the source and effects of managerial bounded rationality on firm scope. That literature, however, does not allow to consider the possibility of transacting ‘outside’ the organization and limits its attention to inside the hierarchy interactions.

<sup>25</sup>This logic provides one way to understand discussions in the business press of the need to ‘focus on the core business,’ outsource the ‘non-core’ assets, etc.

mover advantage: a shared code is suboptimally skewed towards the needs of early adopters; it also shows that there can exist too little code commonality, as, for each group of users, investing in a common code generates positive externalities towards other users.

A final promising research avenue concerns the interaction between organizational codes and labor market dynamics. A worker who learns an organizational code acquires organization-specific human capital. How portable is such capital between organizations? In turn, how does portability affect equilibrium wages and job turnover? Finally, how does the optimal code policy change once the organization realizes that the code it adopts affects the career prospects of its employees and, therefore, its hiring success? Anecdotal evidence suggests that organizations choose very different policies. Some, like Southwest Airlines, strive to imbue their employees with a strong corporate culture that set them apart from the rest of the industry. Other organizations (like university departments and research centers) have an incentive structure that puts a large premium on code portability (to publish, one must communicate with the rest of the profession not just with direct colleagues). Still, others create a distinctive code even though they have high employee turnover (like McKinsey), possibly because they are exploiting a recognized first-mover advantage: their employees are highly valued on the market exactly because they have acquired that particular code.

## References

- [1] Aghion, Philippe and Jean Tirole. “Formal and Real Authority in Organizations” *Journal of Political Economy*, 105 (1): 1—29, 1999.
- [2] Argyres, Nicholas S. “The Impact of Information Technology on Coordination: Evidence from the B-2 ”Stealth” Bomber.” *Organization Science*, 10 (2): 162—180, 1999.
- [3] Arrow, Kenneth J. . *The Limits of Organization*. Norton, New York, 1974.
- [4] Battigalli, Pierpaolo and Giovanni Maggi. “Rigidity, discretion, and the costs of writing contracts.” *American Economic Review* 92(4): 798–817, 2002.
- [5] Blankevoort, P. J. “Contradictory effects of a project management dictionary.” *International Journal of Project Management*, 4 (4): 236—238, 1986.
- [6] Bolton, Patrick and Mathias Dewatripont. “The firm as a communication network.” *Quarterly Journal of Economics*, 104(4): 809–839, 1994.

- [7] Bresnahan Timothy F., Erik Brynjolfsson, and Lorin M. Hitt, “Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence.” *Quarterly Journal of Economics*, 117 (1): 339–376, 2002.
- [8] Brynjolfsson, Eric, and Lorin Hitt. “Beyond Computation: Information Technology, Organizational Transformation and Business Performance.” *Journal of Economic Perspectives*, 14 (4): 23—48, 2000.
- [9] Browning, Larry D., Janice M. Beyer, and Judy C. Shetler. “Building Cooperation in a Competitive Industry: SEMATECH and the Semiconductor Industry.” *The Academy of Management Journal*, 38(1): 113—151, 1995.
- [10] Caroli, Eve and John Van Reenen, “Skill Biased Organizational Change? Evidence from a Panel of British and French Establishments.” *Quarterly Journal of Economics*, 116(4): 1449—1492, November 2001.
- [11] Chatterji, Shurojit and Gragan Filipovich. “Incomplete Contracting due to Ambiguity: Natural Language and Judicial Interpretation.” Mimeo, Colegio de Mexico, 2004.
- [12] Chowdry, Bhagwan and Garmaise, Mark J. “Organization Capital and Intrafirm Communication.” Mimeo, UCLA, 2004.
- [13] Crémer, Jacques. “Corporate Culture: Cognitive Aspects.” *Industrial and Corporate Change*, 3(2): 351—386, 1993.
- [14] Crémer, Jacques. “A Partial Theory of the Optimal Organization of a Bureaucracy.” *Bell Journal of Economics*, 11(2), 683-693, 1980.
- [15] Garicano, Luis. “Hierarchies and the Organization of Knowledge in Production.” *Journal of Political Economy*, 108(5): 874—904, 2000.
- [16] Garicano, Luis and Esteban Rossi-Hansberg. “Organization and Inequality in a Knowledge Economy.” Mimeo, 2003.
- [17] Hart, Paul and Carol Saunders, “Power and Trust: Critical Factors in the Adoption and Use of Electronic Data Interchange.” *Organization Science*, 8 (1), 23—42, 1997.
- [18] Herbold, Robert J., “Inside Microsoft: Balancing Creativity and Discipline.” *Harvard Business Review*, January 1, 2002.
- [19] Herbold, Robert J. “The Fiefdom Syndrom.” Currency Doubleday, New York, New York, 2004.

- [20] Jakob Marschak and Roy Radner. *Economic Theory of Teams*. Yale University Press, New Haven, Connecticut, 1972.
- [21] Melitz, Jacques. Language and Foreign Trade. Mimeo, University of Strathclyde, 2005.
- [22] Prat, Andrea. “Should a Team Be Homogeneous?” *European Economic Review* (46)7: 1187–1207, 2002.
- [23] Rajan, Raghuram G. and Julie Wulf. “The Flattening Firm: Evidence from Panel Data on the Changing Nature of Corporate Hierarchies.”, Working Paper, Wharton School University of Pennsylvania, 2002.
- [24] Roy Radner. “The organization of decentralized information processing.” *Econometrica* 61(5): 1109–1146, 1993.
- [25] Ariel Rubinstein. *Economics and Language*. Cambridge University Press, 2000.
- [26] Claude E. Shannon. “A mathematical theory of communication.” *Bell System Technology Journal* 27: 379–423, 1948.
- [27] Simester, Duncan and Knez, Marc. “Direct and Indirect Bargaining Costs and the Scope of the Firm.” *Journal of Business*, 75(2): 283-304, 2002.
- [28] Sperber, Dan and Lawrence Hirschfeld, “Culture, Cognition, and Evolution,” in Robert Wilson & Frank Keil (eds) *MIT Encyclopedia of the Cognitive Sciences* (Cambridge, Mass. : MIT Press, 1999) pp.cxi-cxxxii.
- [29] Stein, Jeremy. “Information Production and Capital Allocation: Decentralized vs. Hierarchical Firms” *Journal of Finance*, 62(5): 1891-1921, 2002.
- [30] Timothy P. Van Zandt. “Decentralized information processing in the theory of organizations. ” In *Contemporary Economic Issues*, Vol. 4: Economic Design and Behavior (ed. Murat Sertel), MacMillan, London, 1999.
- [31] Wernerfelt, Birger. “Organizational Languages”, mimeo. Available at [http://papers.ssrn.com/sol3/cf\\_dev/AbsByAuth.cfm?per\\_id=26163](http://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=26163), 2003.

## Appendix A

### Proofs of propositions in the text

*Proposition 1. In an optimal code, broader words describe less frequent events: if  $n_k > n_{k'}$ , then  $f_x \leq f_{x'}$  for any  $x \in W_k$  and  $x' \in W_{k'}$ .*

*Proof.* Let  $k$  and  $k'$  be two words such that  $n_k > n_{k'}$  in an optimal code  $\mathcal{C}$ .

>From  $\mathcal{C}$ , construct a new code  $\tilde{\mathcal{C}}$  by moving event  $x$  from word  $k$  to word  $k'$  and event  $x'$  from word  $k'$  to word  $k$ . We must have

$$\begin{aligned} 0 &\geq D(\mathcal{C}) - D(\tilde{\mathcal{C}}) \\ &= d(n_k) p_k + d(n_{k'}) p_{k'} \\ &\quad - d(n_k) (p_k + f_{x'} - f_x) - d(n_{k'}) (p_{k'} + f_x - f_{x'}) \\ &= (d(n_k) - d(n_{k'})) (f_x - f_{x'}), \end{aligned}$$

which proves the result.  $\square$

*Proposition 2. Assume*

$$d(n+1) - d(n) \geq d(n'+1) - d(n') \quad (\text{A.1})$$

*whenever  $n \geq n' \geq 1$ . Unless integer constraints make it impossible, in an optimal code broader words are used less frequently: if  $n_k - n_{k'} \geq 2$ , then  $p_{k'} \geq p_k$ .*

*Proof.* Let  $k$  and  $k'$  be two words such that  $n_k - n_{k'} \geq 2$  in an optimal code  $\mathcal{C}$ .

>From  $\mathcal{C}$ , construct a new code  $\tilde{\mathcal{C}}$  by moving an event  $x$  from word  $k$  to word  $k'$ .

We have

$$\begin{aligned} D(\mathcal{C}) - D(\tilde{\mathcal{C}}) &= d(n_k) p_k + d(n_{k'}) p_{k'} - d(n_k - 1) (p_k - f_x) \\ &\quad - d(n_{k'} + 1) (p_{k'} + f_x) \\ &= [d(n_k) - d(n_{k-1})] p_k - [d(n_{k'} + 1) - d(n_{k'})] p_{k'} \\ &\quad + f_x [d(n_{k-1}) - d(n_{k'+1})] \\ &\geq [d(n_k) - d(n_{k-1})] p_k - [d(n_{k'} + 1) - d(n_{k'})] p_{k'} \\ &\quad (\text{d is increasing}) \\ &\geq [d(n_k) - d(n_{k-1})] (p_k - p_{k'}) \quad (\text{by (2)}). \end{aligned}$$

Because  $D(\mathcal{C}) - D(\tilde{\mathcal{C}}) \leq 0$ , we must have  $p_k \leq p_{k'}$ , which proves the result.  $\square$

*Proposition 3. If  $\tilde{f}$  generates a less complex environment than  $f$ , the minimal diagnosis cost associated with  $\tilde{f}$  is (weakly) smaller than the minimal diagnosis cost associated with  $f$ .*

*Proof.* By proposition 1, we can name the words in the optimal code for distribution  $f$  in such a way that  $k < k'$  implies that  $f_x \leq f_{x'}$  for all  $x \in W_k$  and  $x' \notin W_{k'}$ . This implies

$$d(n_{k-1}) - d(n_k) \geq 0 \text{ for all } k. \quad (\text{A.2})$$

Let  $P_k = \sum_{k' \leq k} p_{k'}$  and  $\tilde{P}_k = \sum_{k' \leq k} \tilde{p}_{k'}$ ; then

$$P_k > \tilde{P}_k \text{ for all } k. \quad (\text{A.3})$$

We have

$$\begin{aligned} \sum_k p_k d(n_k) &= P_1 d(n_1) + (P_2 - P_1) d(n_2) + \dots \\ &\quad + (P_{K-1} - P_{K-2}) d(n_{K-1}) + (1 - P_{K-1}) d(n_K) \\ &= P_1 (d(n_1) - d(n_2)) + \dots \\ &\quad + P_{K-1} (d(n_{K-1}) - d(n_K)) + d(n_K) \\ &\geq \tilde{P}_1 (d(n_1) - d(n_2)) + \dots \\ &\quad + \tilde{P}_{K-1} (d(n_{K-1}) - d(n_K)) + d(n_K) \text{ (by (A.2) and A.3)} \\ &= \sum_k \tilde{p}_k d(n_k). \end{aligned}$$

This concludes the proof, as  $\sum_k \tilde{p}_k d(n_k)$  is not smaller than the minimal diagnosis cost for  $\tilde{p}$ .  $\square$

*Proposition 4. Increasing the number of words from 1 to  $K > 1$  lowers communication costs more for less complex environments. On the other hand, moving from  $K$  words to a very large number of words (perfect communication) lowers communication more for more complex environments.*

*Proof:*

This proposition follows straightforwardly from Proposition 3.

*Proposition 5. The optimal codes contain  $K$  words and satisfy  $\mathcal{C}_A = \mathcal{C}_B$ .*

*Proof.* Clearly, an optimal code saturates the bounded rationality of the engineer, with  $\mathcal{C}_A \cup \mathcal{C}_B$  containing  $K$  words. We must still prove  $\mathcal{C}_A = \mathcal{C}_B$ .

Suppose that  $\mathcal{C}_A \neq \mathcal{C}_B$ , which implies that both  $\mathcal{C}_A$  and  $\mathcal{C}_B$  contain at most  $K - 1$  words. We call  $k^{\min}$  be the narrowest noncommon word of these two codes,<sup>26</sup> and, without loss of generality, assume that it belongs to  $\mathcal{C}_A$ .

<sup>26</sup>That is  $k \in \arg\min_{\tilde{k}} n_{\tilde{k}}$  subject to  $W_{\tilde{k}} \in \mathcal{C}_1 \cup \mathcal{C}_2$  and  $W_{\tilde{k}} \notin \mathcal{C}_1 \cap \mathcal{C}_2$ .

Transform  $\mathcal{C}_B$  into  $\tilde{\mathcal{C}}_B$  by adding  $k^{\min}$  as follows:  $k \in \tilde{\mathcal{C}}_B$  if and only if  $k = k^{\min}$  or  $W = W'/(W' \cap W_k)$  for some  $W' \in \mathcal{C}_B$ . Because  $\#\tilde{\mathcal{C}}_B = \#\mathcal{C}_B + 1 \leq K$ , the bounded rationality of agent  $B$  is still satisfied, and because  $\#(\mathcal{C}_A \cup \tilde{\mathcal{C}}_B) = \#(\mathcal{C}_A \cup \mathcal{C}_B)$ , the bounded rationality of the engineer is also satisfied

For every event  $x \in X$ , the length of the word in  $\tilde{\mathcal{C}}_B$  that contains  $x$  is not larger than the length of the word in  $\mathcal{C}_B$  that contains  $x$ . Moreover, as  $\tilde{\mathcal{C}}_B$  contains one more word than  $\mathcal{C}_B$ , at least one event must be in a strictly narrower word in  $\tilde{\mathcal{C}}_B$  than it was in  $\mathcal{C}_B$ . The new codes are strictly more efficient than the original ones, which proves the result.  $\square$

*Proposition 6. An integrated form becomes relatively more profitable compared to the segregated form when a) the diagnosis cost  $\lambda$  decreases, b) the synergy parameters  $p$  increases, or c) the heterogeneity of the two client distributions  $b$  decreases*

*Proof:*

The formal statement of the proposition is that the ratio  $\frac{\Pi^I}{\Pi^S}$  is increasing in  $p$  and decreasing in  $\lambda$  and  $b$ . Note that this statement implies monotonicity in the choice of the optimal organizational form. Let  $p'' \geq p'$ ,  $\lambda'' \leq \lambda'$ , and  $b'' \leq b'$ . Then, if the optimal form under  $(p', \lambda', b')$  is  $I$ , the optimal form under  $(p'', \lambda'', b'')$  is still  $I$  (because if  $\frac{\Pi^I}{\Pi^S} > 1$  under  $(p', \lambda', b')$ ,  $\frac{\Pi^I}{\Pi^S} > 1$  under  $(p'', \lambda'', b'')$ ). Conversely, if the optimal form under  $(p', \lambda', b')$  is  $S$ , the optimal form under  $(p'', \lambda'', b'')$  is still  $S$ .

The statement that  $\frac{\Pi^I}{\Pi^S}$  is increasing in  $p$  is an immediate consequence of (8). Holding  $\lambda$  and  $b$  constant, the ratio

$$\frac{q^C(p)(1 - \lambda D^*(0))}{q^{NC}(p)(1 - \lambda D^*(b))}$$

is increasing in  $\frac{q^C(p)}{q^{NC}(p)}$ , which in turn is increasing in  $p$ .

Holding  $\lambda$  and  $p$  constant, we have that the ratio  $\frac{\Pi^I}{\Pi^S}$  depends positively on  $D^*(b)$ , which is decreasing on  $b \in (0, 1)$ . Finally, holding  $p$  and  $b$  constant,  $\frac{\Pi^I}{\Pi^S}$  is decreasing in  $\lambda$  because  $D^*(0) > D^*(b)$ .

*Proposition 7. For any  $b$ , if  $p$  and  $\mu$  are high enough, there exist  $1 \leq \lambda_{\min} < \lambda_{\max} \leq 2$  such that the unique optimal organization is*

integrated	if $\lambda < \lambda_{\min}$
hierarchical	if $\lambda \in (\lambda_{\min}, \lambda_{\max})$
separated	if $\lambda > \lambda_{\max}$

*Proof:*

The expected profit in the three organization forms is:

$$\begin{aligned}\pi^I &= q^C(p) (1 - \lambda D^*(0)) \\ \pi^S &= q^{NC}(p) (1 - \lambda D^*(b)) \\ \pi^H &= q^C(p) (1 - \lambda D^*((1 - 2\phi(p))b)) - \mu\end{aligned}$$

Let us begin with the comparison between  $I$  and  $S$ . From the previous proposition, we know that  $\frac{\pi^I}{\pi^S}$  is decreasing (and continuous) in  $\lambda$ . Note also that  $\pi^I > \pi^S$  when  $\lambda = 0$  and  $\pi^S > \pi^I = 0$  when  $\lambda = 2$ . Hence, there exists  $\bar{\lambda}$  such that  $\pi^I > \pi^S$  if and only if  $\lambda < \bar{\lambda}$ .

Next, note that if  $\mu = 0$ , the  $H$  form dominates the  $I$  form for every  $\lambda$  and every  $p$ . In particular, it dominates the  $I$  form for  $\lambda = \bar{\lambda}$ . But given that, by definition,  $\pi^I \geq \pi^S$  when  $\lambda \leq \bar{\lambda}$ , the  $H$  form dominates the  $S$  form for  $\lambda \leq \bar{\lambda}$ .

We have shown that the statement in the proposition holds (with  $\lambda_{\min} < \lambda_{\max}$ ) when  $\mu = 0$ . There exists an interval  $(\lambda_{\min}, \lambda_{\max})$  such that  $I$  is the optimal form when  $\lambda < \lambda_{\min}$ ,  $H$  is the optimal form on  $\lambda \in (\lambda_{\min}, \lambda_{\max})$ , and  $S$  is the optimal form when  $\lambda > \lambda_{\max}$ . By continuity, the statement also holds if  $\mu$  is strictly positive but sufficiently small.

## Appendix B

### Extension of propositions 1 and 2 to the ‘Natural Ordering’ case

Suppose that there is a continuum of events with  $X = [0, 1]$ . The frequency of events is described by a continuous and differentiable, but possibly non-monotonic, probability density  $f$  on  $[0, 1]$ . Words are constrained to be intervals. Writing  $t_0 = 0$  and  $t_K = 1$  a code is therefore a partition<sup>27</sup>  $\{[t_{k-1}, t_k]\}_{k=1, K}$ .

The best K-words code is solution of

$$\min_t \sum_{k=1}^K (F(t_k) - F(t_{k-1})) (t_k - t_{k-1})$$

subject to

$$t_{k-1} \leq t_k \text{ for } k = 1, \dots, K.$$

As in the text, the familiarity of a word,  $[t_k, t_{k+1}]$ , is the probability  $F[t_{k+1}] - F[t_k]$  that the word is used; its breadth,  $t_{k+1} - t_k$ , is the ‘number of events’ in the word. Finally, the average frequency’ of the events in the word is the average density of these events,

$$\phi_k = \frac{F(t_{k+1}) - F(t_k)}{t_{k+1} - t_k}.$$

Then the following proposition contains the results equivalent to propositions 1 and 2 for the case where events are naturally ordered.

**Proposition B.1 (Natural order).** *When words must contain contiguous events, the following two properties hold in an optimal code:*

1. *For two contiguous words, the broader word is used less often .*
2. *For two contiguous words, the broader word describes events which have a lower average frequency.*

We begin by proving the following lemma.

**Lemma B.1.** *In the optimal code  $t_{k-1} < t_k$  for all  $k = 1, \dots, K$ .*

---

<sup>27</sup>As the text is written,  $t_k$  belongs to two words. To avoid this, words should be described by semi-open intervals, at the cost of heavier notation. It should be obvious to the reader that the results are not affected.

*Proof.* Assume for instance that we had  $t_0 < t_1 = t_2 = t < t_3$ . Increase  $t_2$  by a small  $x$ . The diagnosis cost increases by  $\lambda$  multiplied by

$$(F(t+x) - F(t))x + (F(t_3) - F(t+x))(t_3 - t - x).$$

The derivative of this expression with respect to  $x$  for  $x = 0$  is equal to

$$-f(t)(t_3 - t) - (F(t_3) - F(t)) < 0,$$

which proves the result. □

We can now prove the proposition.

*Proof of proposition B.1.* The first-order conditions are

$$F(t_k) - F(t_{k-1}) + f(t_k)(t_k - t_{k-1}) = F(t_{k+1}) - F(t_k) + f(t_k)(t_{k+1} - t_k),$$

which imply

$$f(t_k) = \frac{[F(t_{k+1}) - F(t_k)] - [F(t_k) - F(t_{k-1})]}{(t_k - t_{k-1}) - (t_{k+1} - t_k)} \quad (\text{B.1})$$

The numerator is the difference between the familiarities of contiguous words, while the denominator is the opposite of the difference between their breadths. Thus, optimality requires that the differences between breadth and familiarity of contiguous words have opposite signs, as part 1 of the proposition states.

To prove the second statement, rewrite (B.1) as

$$f(t_k) = \frac{\phi_{k+1}(t_{k+1} - t_k) - \phi_k(t_k - t_{k-1})}{(t_k - t_{k-1}) - (t_{k+1} - t_k)}$$

Thus  $\phi_{k+1} - \phi_k$  and  $(t_{k+1} - t_k) - (t_k - t_{k-1})$  must be of opposite sign: that is, events in the broader word have a lower average frequency. □