Language and the Theory of the Firm[1]

Jacques Crémer
Université de Toulouse, IDEI-GREMAQ and CNRS

Luis Garicano
University of Chicago and CEPR

Andrea Prat
London School of Economics and CEPR

March 15, 2006

[1]A previous version of this paper circulated under the title: 'Codes in Organization.' We thank Philippe Aghion, Karen Bernhardt, Ellynne Dec, 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 Workhop, and at various universities for useful comments and Pedro Vicente for outstanding research assistance. Garicano thanks the Toulouse Network on Information Technology for financial support.

Abstract

An organization will often use a specialized technical language that is understood by its members but not by others. 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" are suboptimal: two organizations will either share the same language or develop two entirely distinct set of technical words. This endogenous discontinuity in communication structure results in a discontinuity in firm structure, and a limit on firm scope. A broader scope allows for more synergies to be captured, but at the cost of less precise communication within each unit. Our theory reconciles two recent phenomena within organizations: the recent increase in information centralization and the reduction in hierarchical centralization.

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. These "codes," to use the term introduced by Arrow (1974), economize in communication costs, as they allow for 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 trade-offs 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.

Coordination failures due to 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. Investigators found no individual guilty; the tragedy was the result of grave organizational dysfunctions.[1] In particular, communication was hindered by misunderstandings resulting from the different, and incompatible, codes used by the Army and the Airforce. Three instances are particularly striking. First, the word 'aircraft' 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

[1] The account that follows is from Snook (2000).

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). Third, the Air Force and the Army helicopters interpreted differently 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'), which led the Army helicopters to be 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.

For a less belligerent example, consider the development of a common language for DNA sequencing at the world's foremost genomics research center, the Broad Institute in Cambridge, MA.[2] Originally the Institute was organized along functional lines (e.g. molecular biology, production, technology development, production informatics). While communication within each group was satisfactory, between group communication was poor: 'scientists and engineers from different areas could not understand each other's language.' Yet, obtaining further process improvements required multidisciplinary insights. In 2002-2003, the institute decided to develop a common language based on the language of statistical control systems, including common concepts such as process risk, Pareto categorization of risks, process variability, design phase curve, etc. This common language allowed for a highly successful organizational re-design around multi-disciplinary groups for large projects. Not only did coordination between the different groups improve, but also managers were able to reduce the time they spent on coordination, consistently with the theory we present in this paper.

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) on the use of specialized codes by organizations. We then use this theory to understand the constraints on firm scope and structure imposed by the need for specialized languages, and to derive testable empirical implications.

Section 2 studies a simple model of codes. We begin by studying communication between on agent who receives a stream of heterogeneous problems (e.g., customers with certain needs) and must communicate this information to someone else in the organization who holds specialized knowledge (e.g.a production engineer). As the two agents are boundedly rational, they can only learn a limited number of words to describe the characteristics of the

[2]Personal interview of one of the authors with Robert Nicol, director, Sequencing Operations, Broad Institute.

problems they face. Hence, the problem-solving capability of an organization depends on the code it uses. Following Arrow, we make a key assumption: the organization is able to determine the code its members adopt.[3]

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.

After studying the optimal code when two agents are communicating, we turn to the situation where two agents receive problems and must communicate them to one agent. We show that bounded rationality imposes sharply decreasing returns to the diversity of codes: groups of homogeneously skilled agents will use either entirely separate codes or common codes: "dialects," as we show, cannot be optimal. This code commonality is a key determinant of the decreasing returns to scope in organizations. Thus, it shapes both the scope of organizations 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 highlights the trade-off between improving local efficiency and generating synergies. Broadening the scope of a firm is justified since it can exploit some synergy. However, firms cannot capture the synergy and leave everything else the same. Capturing it requires some coordination and communication between services, which requires in turn that they speak a common language. Since a common language cannot be well-suited to the needs of diverse specialized individual services, the scope of the firm is limited. Bounded rationality thus results in a theory of firm boundaries. We identify the variables that determine the terms of this trade-off. A broader firm, which must use 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 — in these cases the optimal codes are close and the distortion required by the common code is small. In its study of this tradeoff, our paper is a contribution to the debate on the tradeoff between specialization and coordination (Hart and Holmstrom (2002), Hart and Moore (2005)), but from an entirely different angle (endogenous communication).[4] Second,

[3]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. Section 5 discusses the connection to Rubinstein's work.

[4]Hart and Holmstrom (2002) study the trade-off betwen the synergies obtained from

codes interact in a 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.

Our analysis sheds some light on the impact of information technology-related changes in search cost on organization. As these costs decrease, we expect to observe horizontal integration, since the cost of common codes decrease and capturing synergy becomes relatively more attractive. We also expect to observe vertical disintegration in the form of 'delayering,' 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 and horizontal coordination, a reduction in the number of layers of management within existing hierarchies 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, together with 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. As we argue in Section 4, although one could expect, absent our theory, that more centrally available information would lead to more central decisions, it appears that the opposite is true, as we document in our case studies. Our analysis suggests that what is crucial is the homogenization 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 examples of these changes: the Microsoft corporation, where common human resource and finance categories were adopted throughout the organization, leading, according to the account

integration and the private benefits of control obtained by managers. Hart and Moore (2005) study the allocation of authority over the use of assets when agents with 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 organization's. The key comparative static prediction in their work concerns the consequences of synergies on integration and hierarchy. Our analysis generates similar comparative statics on integration and on hierarchy (where hierarchy is unrelated to authority) but, because communication between agents is possible in our work but not in theirs, also allows us to evaluate the impact of information costs on horizontal and vertical structure.

of Robert J. Herbold, its Chief Operating Officer 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 replaced by horizontal, peer-to-peer communication thanks to a common set of categories across teams.

Our paper abstracts from incentive considerations. This useful simplication is common in organizational theory.[5] However, to the best of our knowledge, none of the previous literature has studied the relationship between the organizational code and the organizational choices of the firm. In this respect, our analysis is the first to provide a simple way to analyze an elusive idea, the idea of organizational language, and to use such a formalization to study organizational isses.[6]

Section 5 discusses links with 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.[7]

[5] An incomplete list includes papers on team theory (Marschak and Radner's (1972), Crémer (1980)), information processing (Radner (1993) and others: see Van Zandt, 1999 for a survey), problem solving (Garicano, 2000), communication within organizations (Bolton and Dewatripont (1994)), and corporate culture (Crémer (1993), Prat (2002), Chowdhry and Garmaise (2004)).

[6] Wernerfelt (2004) considers codes that minimize communication costs under common preferences, and studies the existence of symmetric or asymmetric equilibria with one or multiple codes; he does not consider the interaction between codes and organization. An alternative approach, which studies endogenous communication in a strategic environment, is taken e.g. by Battigalli and Maggi (2002) who use a sophisticated model of language to develop a theory of contract incompleteness; and by Dewatripont and Tirole (2005) who study the strategic interactions between the communication efforts made by different agents.

[7] 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 the representation the messages with sequences of binary digits (bits) that are as short as possible and understanding issues such as the capacity of different types of channels or the determinants of decoding errors. 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

2.1 Model

A team composed of a salesman and an engineer serves clients. The clients approach the salesman with a problem that demands a solution x, drawn with probability $f_x > 0$ from a finite set X.[8] The salesman can classify problems, but not perfectly, since he is boundedly rational. The engineer and the salesman therefore agree, before the salesman meets the client, on a code – a shared technical language that allows the salesman to transmit (coarsely) the class to which he has assigned the client's problem. Formally, a code $\mathcal{C}$ is a partition $\{W_1, W_2, \ldots, W_K\}$ 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 call the *breadth* of word k the number $n_k \equiv \sharp W_k$ of events that it contains, and 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, the engineer must still identify the precise problem x of the client in order to solve it. This *diagnosis* stage takes time and/or energy, and its cost is a strictly increasing function d of the breadth n_k of word k — the less precise the word, the more work diagnosis requires. The expected diagnosis cost associated with code $\mathcal{C}$ is therefore

$$D(\mathcal{C}; f) = \sum_{k=1}^{K} p_k d\left(n_k\right), \tag{1}$$

which we will simply write $D(\mathcal{C})$ when there is no risk of ambiguity.[9]

We shall assume that serving clients is sufficiently valuable relative to the diagnosis cost that the engineer would never want to exclude solving problems with certain values of x. Thus 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

implications of agents' bounded rationality, the transmission cost is given, but the sender is prevented from transmitting all the information.

[8] Nothing fundamental in our analysis requires a finite set, but many technicalities are avoided by this assumption. Similarly, to avoid unnecesary technilaties, we assume that $f_x > 0$ for all events x.

[9] Linear diagnosis costs are a special case, and we use them because they provide a tractable technique to study the organizational consequences of codes. Moreover, some models of diagnosis do yield linear costs, such as when all objects must be searched for example to be compared. In any case, if search costs are low compared to the benefits of a better fit, searching through the entire set of objets would be (close to) optimal and the analysis of the text carries through.

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.[10]

Throughout this section, the frequency f_x is exogenous. In reality, the frequency of events is influenced by the organizational design — we explore this link in Section 3.

2.2 Optimal codes

In this subsection, we derive some properties of the optimal code for these two agents (remember that all proofs are in the appendix).

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 in an optimal code events of similar frequencies are grouped together: 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.

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

Proposition 2. *If the function d is "convex" in the number of events n, i.e., if*

$$d(n+1) - d(n) \geq d(n'+1) - d(n') \text{ for all } n \geq n' \geq 1,$$

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

[10] 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. These experiments allow Munyan and Camerer (2005) to study mergers and find that the merged groups eventually reach the performance of the non-merged groups.

The intuition for the result can be easily seen by assuming $n_k \geq n_{k'} + 2$ and $p_k > p_{k'}$, and considering the special case where there exists an event $\tilde{x} \in W_{k'}$ such that $p_k - p_{\tilde{x}} > p_{k'} + p_{\tilde{x}}$. Then, transferring $\tilde{x}$ from W_k to $W_{k'}$ decreases the diagnosis cost of the most probable word, W_k, at least as much as it increases the cost of the less probable word, $W_{k'}$, which is impossible if the code is optimal.

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, 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

Agents face environments with varying degrees of complexity, and this affects the value of the codes that they use. We identify here complexity with the variability of the environment – a more complex environment is a less predictable one, where a wider range of problems is likely to be confronted. We illustrate this fact with a simple example. Suppose there are three possible events $\{x_1, x_2, x_3\}$ and that $\Pr[x_3] = p$ and $\Pr[x_1] = \Pr[x_2] = \frac{1}{2}(1-p)$, with $p > 1/3$. If the agents can use at most two words, the optimal language consists of a word for $\{x_1, x_2\}$ and a word for $\{x_3\}$. The expected diagnosis cost is $p + 2(1-p) = 2 - p$. Diagnosis cost is decreasing in the probability of the most likely event, p. As we shall see, this is a general property. The diagnosis cost always goes down when we move from a more complex environment (one where it is hard to predict what event will occur) to a simpler one.

Consider two distributions of the same set of events, f and $\tilde{f}$. Use the first distribution, f, to rank events according to their probability of occurring, namely, $f_1 \leq f_2 \leq ... \leq f_m$. Let F_i denote the probability that the event has index i or lower given distribution f. Similarly, let $\tilde{F}_i$ denote the probability that the event has index i or lower given distribution $\tilde{f}$ (but recall that we are still using the indexing derived using f). We then say:[11]

Definition 1. *The distribution f represents a more complex environment than $\tilde{f}$ if $F_i \geq \tilde{F}_i$ for all events i.*

[11] Our notion of environment simplicity is analogous in spirit to first-order stochastic dominance in that it induces a partial ordering of distribution functions.

An environment becomes simpler if unlikely events become even less likely and likely events become even more likely. In the three-event example above, for any $\tilde{p} > p > 1/3$, the distribution with p represents a more complex environment than the distribution with $\tilde{p}$.

We can show that the diagnosis cost is increasing in complexity:

Proposition 3. *If f represents a more complex environment than $\tilde{f}$, the minimal diagnosis cost associated with f is (weakly) larger than the minimal diagnosis cost associated with $\tilde{f}$, that is* $\min_{\mathcal{C}} D(\mathcal{C}; f) \geq \min_{\mathcal{C}} D(\mathcal{C}; \tilde{f})$.

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 diagnosis cost.

When the number of words is equal either to 1 or to the number of events, diagnosis costs are the same whatever the complexity of the environment. Together with proposition 3, this implies an interesting interaction between the benefit of additional words and the complexity of the environment.

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

Having a code with more words is always useful, but its relative benefit depends on both the complexity of the environment and the richness of the language. Starting from the coarsest language, adding words is most beneficial in simple environments. The savings in terms of diagnosis cost are high because few words can describe precisely a large proportion of the events. If instead the code is already rich, the additional cost reduction is greater for complex environments. As a consequence, if words are expensive, we should observe basic codes for simple environments but no code at all for complex ones. With lower cost of words, the code will stay basic in simple environments, but could jump from non-existing to rich in complex ones.

2.4 Shared codes and dialects

We study now the choice of code when agents facing different distributions of events must communicate with one another. Specifically, we study the choice

of code by an engineer who needs to communicate with two[12] salesmen, A and B, who face the same set of events X but different distributions f_x^A and f_x^B. We shall show that, in the stark model which we are analyzing, "dialects" are never optimal. Each 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, must know code $\mathcal{C}_A \cup \mathcal{C}_B$ (of course, he only uses the relevant part of this code when communicating with either salesman).

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

$$\mathcal{C}_A = \{\{1, 4\}, \{2, 5\}, \{3, 6\}\}, \tag{2}$$
$$\mathcal{C}_B = \{\{1, 2, 3\}, \{4, 5, 6\}\}. \tag{3}$$

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 $\mathcal{C}_A = \mathcal{C}_B$.*

The proof of proposition 5 can be found in the appendix, but we present two examples which illustrate it. First, with $K = 5$, consider the codes of equations (2) and (3). The narrowest noncommon words[13] 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 $\tilde{\mathcal{C}}_B$, we obtain the code composed of ($\{1, 4\}$, $\{2, 5\}$, $\{3\}$, $\{6\}$ and $\{3, 6\}\}$). This code satisfies the bounded rationality of the agents, and its use by all will lead to lower diagnosis cost than $\mathcal{C}_A$ and $\mathcal{C}_B$.

A more complicated example starts from

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

If we take $\{1, 2, 3\}$ as the narrowest non-common word, we obtain

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

[12] 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.

[13] 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.

All events but 10, 14, 15 and 16 are now represented by strictly shorter words and no event is represented by a longer word. It is more efficient for the engineer to use the code $\mathcal{C}_A$ and $\tilde{\mathcal{C}}_B$ than the initial codes $\mathcal{C}_A$ and $\mathcal{C}_B$. Iterating the elimination of the narrowest non-common word, we converge to a common code for both agents which is more efficient than both of the original codes.

We have shown that the engineer will use the same code to speak to both salesmen. Which code will be chosen? 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 following corollary.

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 finite 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 select for the role of engineer the agent who is able to learn the greatest number of words. 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. On the other hand, a more realistic model would allow for communication between the two salesmen, which would reinforce the benefits of a common language. Garicano and Rossi-Hansberg (2005) build a theory 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 Section 3.3, where we assume that the firm can hire, at an additional cost, an agent with a larger K.

Thus while our result that there are no dialects, i.e., that the language used for all communications is exactly the same, is "too strong," we believe that it draws attention to a general phenomenon of great economic significance: namely, the steeply decreasing returns to dialect variety. Suppose that an agent develops different specialized words to talk with different subsets of agents. Since the union of the words in all of these dialects must satisfy each agent's bounded rationality constraint, each dialect can only contain a

limited vocabulary. If the agent instead uses a general common code to talk to all other agents, he can have a less tailored, but richer set of words that he can use in his communication with each of the other agents. The common language forces agents to give up on tailoring words to specific needs, but relaxes the constraints imposed by bounded rationality. Alternatively, agents may choose fully separate languages where they can enjoy the advantage of tailoring each word to their specific needs without giving up on bounded rationality.

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. Given an environment, what are the optimal communication structure and code? And what is the organizational structure that supports them?

We develop a simple model with two services. We compare three possible organizational forms: *separation*, where the two services use different codes; *integration* where the two services share the same code; *translation* or *hierarchy* where there exists a hierarchical structure supplying an interface between the services. The three forms are presented schematically in Figure 1 and described next. We study communication and coordination in these three forms, and we identify the environments in which each of them is optimal.

3.1 The costs and benefits of collaboration

Services A and B are each composed of one salesman and one engineer. They generate revenue, normalized to 1, whenever an engineer correctly diagnoses the problem faced by the client of a salesman. A fraction of clients ν_A from the overall client population arrive at service A, and a fraction ν_B arrive at service B, with $\nu_A+\nu_B = 1$. We call f_A and f_B the problem distributions over these subpopulations, and f the distribution over the entire client population, that is $f = \nu_A f_A + \nu_B f_B$. We call the quadruple (f_A, f_B, ν_A, ν_B) a *client distribution*.

Collaboration is beneficial because there exist some synergies between the services. In particular, if the two services A and B do not collaborate with each other (Panel A in Figure 1) they can deal with a number q^{NC} (NC stands for "No Communication") of clients on average: service A can process $\nu_A q^{NC}$ while B can process $\nu_B q^{NC}$clients. If instead the two services can cooperate

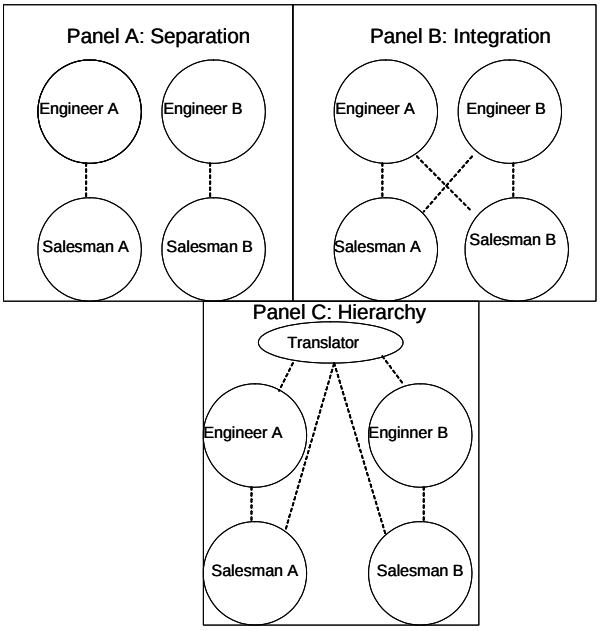

Figure 1: Communication in three possible organizational forms. The dashed lines represent lines of communication.

by serving some of each other's clients (which requires communication, see Panel B in Figure 1), they can serve a higher number q^C of clients, $q^C > q^{NC}$: $\nu_A q^C$ are clients of salesman A and $\nu_B q^C$ of salesman B. However, these clients will be served in part by the other engineer: for instance, $\phi\nu_A q^C$ clients of salesman A will be served by engineer A while $(1-\phi)\nu_A q^C$ will be served by engineer B. The following example provides a very simple situation that falls under these general assumptions.

Example 1. Assume that an engineer has the ability to attend to the needs of at most one client per period, and that the number of clients who contact a salesman in each period is

$$\begin{array}{ll} 0 & \text{with probability } p, \\ 1 & \text{with probability } (1-2p), \\ 2 & \text{with probability } p, \end{array}$$

where $p \in [0, 1/2]$. Then

$$q^{NC} = 2\left((1-2p)+p\right) = 2\left(1-p\right).$$

If they collaborate, the firm can divert business from an overburdened service to an unemployed one. When one salesman

receives zero clients and the other one receives two, which happens with probability $2p^2$, the client is served by the engineer of the salesman with no client. Then,

$$q^C = 2\left(1 - p + p^2\right),$$

with the clients served by the other engineer are determined by:

$$\phi = \frac{1-p}{1-p+p^2}.$$

The aggregate profits of the two services are given by the value of the clients serviced minus the per client diagnosis cost. We let $\lambda D^*(f)$ be the output cost of diagnosis associated with a distribution of events f when the optimal code is chosen, that is

$$\lambda D^*(f) = \min_{\mathcal{C}} \lambda D(\mathcal{C}; f)$$

where λ, the 'diagnosis cost' parameter expresses diagnosis cost in output units.[14]

Then we have that the profits of the two services when non-integrated are given by

$$\Pi^S = q^{NC}\Big[1 - \lambda(\nu_A D^*(f_A) + \nu_B D^*(f_B))\Big]. \tag{4}$$

When services do collaborate, salesmen need to be able to communicate customer needs to both engineers. 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 $\mathcal{C}^*(\nu_A f_A + \nu_B f_B)$, the language which would be optimal with one engineer

[14] We assume throughout that it is profitable to use a salesman, so that the engineer does not benefit from dealing with clients that have not been first dealt with by a salesman. If information did not transit through a salesman, the engineer would have diagnose the problem on an interval of size 1, at cost λN_X, where N_X is the number of events in X; the net profit per client would be $1 - \lambda N_X$. Thus technically, we assume

$$\lambda > \lambda^{\min} \stackrel{\text{def}}{=} \frac{1}{d\left(N_X\right)},$$

so that it is profitable to use a salesman. Similarly, as shown in 2.3, the diagnosis cost is highest when the distribution of types is the uniform distribution f^{unif}; therefore, $q^{NC}[1 - \lambda D^*(f^{\text{unif}})]$ is a lower bound on profits when using a salesman. We assume

$$\lambda < \lambda^{\max} \stackrel{\text{def}}{=} \frac{1}{1/D^*(f^{\text{unif}})}.$$

receiving messages from one salesman for which the distribution of characteristics is the distribution of characteristics of the client population. Thus total profits of collaborating services are

$$\Pi^I = q^C \Big[1 - \lambda D^*(\nu_A f_A + \nu_B f_B) \Big]. \tag{5}$$

In what follows, we will want to study the determinants of organizational form by analyzing how the difference between integration and separation (Π^I -Π^S) depends on the environment. One important determinant of the cost of a merger is the difference between the pre-merger optimal codes and the post-merger common code; this in turn depends on how similar the distributions of problems in the two distributions of clients are. We define the homogeneity of the two client distributions formally next.[15]

Definition 2. *Given a client population f, the client distribution among services $(\tilde{f}_A, \tilde{f}_B, \tilde{\nu}_A, \tilde{\nu}_B)$ is more* ***homogeneous*** *than the client distribution (f_A, f_B, ν_A, ν_B) if the distribution over the entire population f is the same under both and the distributions of types $\tilde{f}_A$ and $\tilde{f}_B$ are convex combinations of f_A and f_B.*[16]

3.2 Integration or Separation?

In this subsection, we study the choice between integrated and separated services focusing on the comparative statics of that choice. Intuitively, it is clear that the key loss due to collaboration results from the deterioration in within-service communication when a less specialized language is used. We show this next.

The fundamental result on which our analysis rests is the concavity of the diagnosis cost.

Lemma 1. *For any two distributions f_A and f_B and any $\alpha \in [0,1]$, we have*

$$D^*(\alpha f_A + (1-\alpha) f_B) \geq \alpha D^*(f_A) + (1-\alpha) D^*(f_B).$$

A consequence of Lemma 1 is that it is easier for an organization to treat different client populations separately by developing a separate specialized

[15] There is no immediate connection between the notion of heterogeneity of a client distribution and the notion of environment complexity used earlier. In particular, the fact that $(\tilde{f}^A, \tilde{f}^B, \tilde{\nu}_A, \tilde{\nu}_B)$ is less heterogenous than (f^A, f^B, ν_A, ν_B) does not imply that $\tilde{f}^A$ is more complex than f^A or that $\tilde{f}^B$ is more complex than f^B.

[16] For an algebraic expression of the conditions in this definition, see equations (A.2) and (A.3) in appendix A.

code for each one of them.[17] This result is critical for the comparison between separate and integrated organizations. Inspecting (4) and (5) , it is clear that while synergies make integration profitable, $q^C > q^{NC}$, a merger results in a deterioration in communication within both units, as (recall $\nu_B = 1 - \nu_A$) : $(1 - \lambda D^*(\nu_A f_A + \nu_B f_B)) \leq (1 - \lambda(D^*(f_A) + \nu_B D^*(f_B)))$. Thus, depending on the values of the parameters, the profit under integration can be greater than, equal to, or smaller than the profit under separation. The following proposition describes the comparative statics characterizing the choice of organizational form.

Proposition 6. *An integrated form becomes relatively more profitable compared to the segregated form as a) the diagnosis cost λ decreases, b) the synergy between the two services, measured by q^C/q^{NC}, increases, or c) the homogeneity of the two client distributions increases.*

Parts a) and b) are easily proven by examination of (4) and (5); they are also very 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. Similarly, if synergies are high, an integrated form allows for larger increases in profitability. Part c (homogeneity), although less straightforward to prove, is also intuitive. The separate codes in more homogeneous populations are not particularly well adapted to each individual population, and thus achieve relatively high diagnosis costs. This implies a lower loss in the case of merger.

Note that we measure synergy by the ratio of number of customers served when the services are cooperating to the number when they are not. This formulation is agnostic as to whether an increase in synergy is due to a generally positive event ($q_1^C > q^C$ and $q_1^{NC} > q^{NC}$), a generally negative event ($q_1^C < q^C$ and $q_1^{NC} < q^{NC}$), or one that is positive only for communicating services. A higher level of synergy between services will clearly make it more profitable to have the two services communicate.

Example 1 (cont.) In our example, we have

$$q^C/q^{NC} = \frac{2(1-p+p^2)}{2(1-p)} = 1 + \frac{p^2}{1-p};$$

[17] The proof of Lemma 1 is unchanged if the cost of using a word were to depend on the specific events in the word and not only on their number — for instance, the word $\{1, 2\}$ could be cheaper to use that the world $\{1, 4\}$ as 1 and 2 are closer than 1 and 4. Since the lemma is the basis for all the results that follow, the fact that the lemma still holds in a such a more general model means these results would still be valid under this more general hypothesis.

an increase in p, which is generally a negative event, does lead to an increase in synergy.

3.3 Hierarchy

We consider now organizational structures that allow the two services to use different codes, but exploit the synergy by employing a fifth agent who provides translation (see Panel C in Figure 1) and is hired at cost μ. That is, rather than having all agents use the same code to allow for horizontal (between services) collaboration, the translator steps in when inter-service communication is needed.[18] If salesman A has a problem to transmit to engineer B, he communicates to the translator the type of the problem in the code used in service A. The translator will search for x, and then he will transmit the information to engineer B in the code used in service B.

Although the mode of transmission of information is different in the hierarchical and in the integrated form, the mode of collaboration is the same - clients can be reallocated through the intervention of the translator. As a consequence, engineer A will serve $\nu_A\phi q^C$ clients of salesman A and $\nu_B(1-\phi)q^C$ clients of salesman B. Hence, the language that the translator will use to speak to him will be the language appropriate for the distribution

$$\frac{\nu_A\phi}{\nu_A\phi+\nu_B(1-\phi)}f_A+\frac{\nu_B(1-\phi)}{\nu_A\phi+\nu_B(1-\phi)}f_B.$$

The profits from the hierarchical organization will therefore be

$$\Pi^H=q^C\Bigg[1-\lambda\Bigg[(\nu_A\phi+\nu_B(1-\phi))D^*\left(\frac{\nu_A\phi f_A+\nu_B(1-\phi)f_B}{\nu_A\phi+\nu_B(1-\phi)}\right)+(\nu_A(1-\phi)+\nu_B\phi)D^*\left(\frac{\nu_A(1-\phi)f_A+\nu_B\phi f_B}{\nu_A(1-\phi)+\nu_B\phi}\right)\Bigg]\Bigg]-\mu. \quad (6)$$

Obviously, the distribution of clients in a hierarchical organization is less homogeneous, in the sense of definition 2, than the distribution in the integrated organization, and more homogeneous than the distribution in the separated organization. Therefore, the diagnosis cost is intermediate between the average diagnosis cost in these two other organizations.

[18] We assume that the translator has no direct diagnosis cost (his $\lambda=0$). We obtain the same qualitative results if we assume that the the translator faces a direct diagnosis cost which is lower than the other agents' cost. The assumption that the translator has a lower cost is natural, given that he specializes in communication.

We can now generalize and extend the result in proposition 6 to the comparison between the three organizational forms. We assume that the separated form does not dominate the integrated form for all values of λ.[19]

Proposition 7. *There exists $\mu^* > 0$ such that, for all $\mu \in (0, \mu^*)$, there exists $\lambda^{IH}(\mu)$ and $\lambda^{HS}(\mu)$ such that the unique optimal organization is*

$$\begin{array}{ll} \textit{integrated} & \textit{if } \lambda < \lambda^{IH}(\mu), \\ \textit{hierarchical} & \textit{if } \lambda \in \left(\lambda^{IH}(\mu), \lambda^{HS}(\mu)\right), \\ \textit{separated} & \textit{if } \lambda > \lambda^{HS}(\mu). \end{array}$$

The function λ^{IH} is decreasing, while λ^{HS} is increasing.

Translation induces a fixed cost μ and increased diagnosis costs, but makes inter-service communication possible and thus allows the services to profit from the existing synergies. Since translation allows the two services to use efficient service-specific codes rather than a common code, it will be preferred to integration when well adjusted codes are more important, that is when λ is large. However, if the diagnosis cost λ is too high, then the synergy gains do not compensate the increased communication costs, and the optimal structure is the separated form. Figure 2 illustrates these effects (see Example 1, below for the description of the underlying model in the example of this figure).

> Example 1. (cont). The externality is as above, and half of the clients arrive at each salesman's. For both services we have $X = \{1, \ldots, N\}$, with $f_x^A = (3N - 2 + 4x)/(5N^2)$ and $f_x^B = (7N + 2 - 4x)/(5N^2)$. This implies that f_x^A is increasing in x while f_x^B is decreasing. Because $(f_x^A + f_x^B)/2 = 1$ the aggregate client distribution is uniform. The agents can use only two words, and one can show that when N becomes large, when the services are separate, for each of them the approximately 55% events with the smallest probability are allocated in one word, while the others are allocated to the other word. Figure 2 shows the optimal organizational forms when N is very large.

Intuitively, Proposition 7 implies that the two modes of between service coordination in our model, a common code and hierarchy, are substitutes. An organization may attempt to capture synergies in two ways: either by

[19] For a low μ, it may be the case that $\lambda^{HS}(\mu) = \lambda_{\max}$. However, for μ close to μ^*, it must be that case that $\lambda^{HS}(\mu) < \lambda_{\max}$ and $\lambda^{IH}(\mu) > \lambda_{\min}$. See the proof in the appendix for the definition of μ^*.

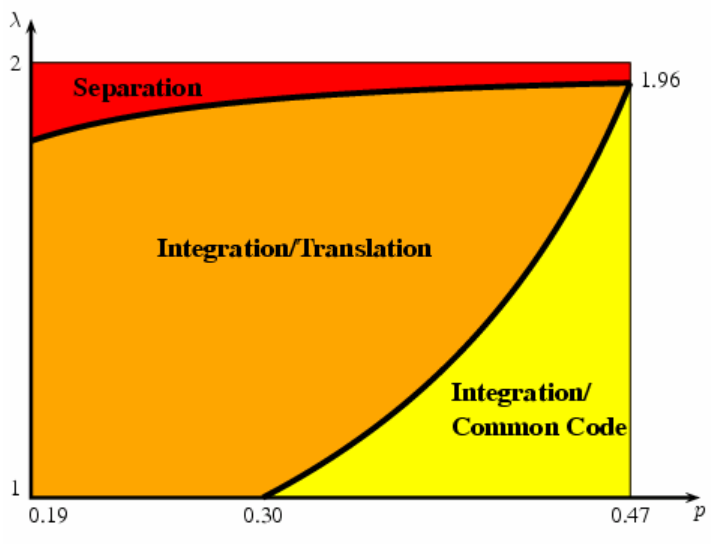

Figure 2: The choice between separation with or without a hierarchy for example 1, as a function of the cost of communication λ and of the synergy parameter p.

having a manager who translates the needs of one service for the other; or through a common language that allows the members of the two services to communicate with each other. A common language will be preferred when either the costs of imprecise communication are low, or when synergies are sufficiently high.

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 while decision-making becomes more decentralized. Such changes are paradoxical from an agency perspective, since increasing information centralization should go hand in hand with centralization of decision rights. Agency costs are tolerated in order 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. Our theory, through the substitution of hierarchy by common codes shown in Proposition 7, makes the opposite prediction. 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.[20]

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 it, of accessing it when it is contained in databases, and of processing it. For the purpose of our analysis, a key feature of these changes is that by reducing the cost of searching information, these advances reduce the cost of receiving an imprecise message from a third party, as the receiver can conduct the follow up diagnosis of the problem more cheaply. 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 — we should witness the substitution of common codes for hierarchies and increase in peer-to-peer communication. As we have argued in the introduction, these predictions are specific to our model, since it is the first in the literature that allows communication for both horizontal (between similar agents) and vertical agents. Both of these broad predictions appear consistent with the evidence.[21]

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.[22] As information

[20] 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 trade between countries. See Melitz (2005) for evidence and references to previous work.

[21] Garicano and Rossi-Hansberg (2006) which studies the organization of knowledge acquisition and earnings inequality, show how 'delayering' can also result from a decrease in the cost of accessing knowledge, which leads to an increase in the knowledge available to each individual, and thus limits his reliance on the hierarchy for problem solving.

[22] 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

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[23] and, earlier, Electronic Data Exchanges (EDI)[24]. These programs allow for the exchange of electronic data by standardizing its format. Through these systems, firms have substituted flexible ways to code their data by more rigid but unified central databases.[25] These common information systems have resulted in increasing horizontal information links within and between firms, as the examples below show.[26]

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 way that our model describes. We use some case-study evidence to illuminate this connection.

---

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.)

[23] See for instance the products offered by SAP `http://sap.com` or Baan `http://www.baan.com`.

[24] We refer to EDI systems broadly, to include other related approaches such as CPFR ("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 Webmans' 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).

[25] 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).

[26] As we will discuss in the conclusions, more work is needed on whether the horizontal scope of organizations has been increased by IT.

Microsoft. Robert J. Herbold,[27] 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 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."

The German country manager provides another example:

> "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."[28]

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.

[27] We rely on Herbold personal account, in his *Harvard Business Review* article. All the quotes below are from his account.

[28] 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.

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.[29] Advances in information technology made it possible for Northrop, Boeing, Vaught (a division of LTV) and General Electric, the 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."[30] 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 hierarchical coordination, since among the main consequences of the creation of a relatively rigid unifying code 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 Conclusions

By formalizing Arrow's idea 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 between the choice of organizational language and the choice of organizational structure. This has led to testable implications on how changes in processing costs

[29] The account that follows draws 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`.

[30] "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).

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.

Our work is related to Rubinstein's (2000) work on the economics of language. In sections 1.3 and 1.4 of his boook, he proposes a model of optimal languages, although in a very different framework (see also Econometrica paper). A language is a binary relationship between "events" (in our language) and the optimal language allows for an efficient precise identification of specific events. Rubinstein also studies the emergence of languages through evolution (chapter 2); his analysis provides some justification for focusing on efficient languages, as we do, as he shows that they are selected in the evolutionary model that he considers. Of course, although we do share Rubinstein's view that the study of optimal languages is important and fruitful, our main aim in this paper is different, as we are mostly interested in the interplay between language and organization.

Our analysis suggests several interesting avenues for new research. First, our model yields testable hypotheses. The availability of large databases of business texts and their ease of access may allow for a study of the determinants of the commonality of the language used across different services of different firms or across different firms in an industry. Such research would also allow for a direct test of our hypothesis on the substitution of codes for hierarchies: one should observe more 'de-layering' (less hierarchy) and more horizontal communication as codes become more common.

Second, 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 down the chain of command or in space. The exact meaning of the word, the exact event within it that is referred to, is soft information: it is not too costly to figure out this meaning in one to one communication, but it is very costly to do so along a chain of command or across a long distance.

Third, 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 choose codes independently and act strategically. A previous version of this paper, available from the authors, studies such choices. Our analysis identifies a first 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. Argyres' study of the B-52 bomber provides some illustration of this phenomenon, as the code which was finally adopted was very similar to the code already used by Northrop.

Finally, it would be useful to analyze 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 Andersen Consulting, strive for uniqueness, while others, like university departments and research centers put a large premium on code portability (to publish, one must communicate with the rest of the profession,not just with direct colleagues).

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] Dewatripont, Mathias and Jean Tirole. "Modes of Communication", 113 (6), 2005, p. 1217-1238.

[16] Garicano, Luis. "Hierarchies and the Organization of Knowledge in Production." *Journal of Political Economy*, 108(5): 874—904, 2000.

[17] Garicano, Luis and Esteban Rossi-Hansberg. "Organization and Inequality in a Knowledge Economy." *Quarterly Journal of Economics.* Forthcoming, 2006.

[18] 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.

[19] Herbold, Robert J., "Inside Microsoft: Balancing Creativity and Discipline." *Harvard Business Review*, January 1, 2002.

[20] Herbold, Robert J. "The Fiefdom Syndrom." Currency Doubleday, New York, New York, 2004.

[21] Jakob Marschak and Roy Radner. *Economic Theory of Teams*. Yale University Press, New Haven, Connecticut, 1972.

[22] Melitz, Jacques. Language and Foreign Trade. Mimeo, Univeristy of Strathclyde, 2005.

[23] Prat, Andrea. "Should a Team Be Homogeneous?" *European Economic Review* (46)7: 1187–1207, 2002.

[24] 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.

[25] Radner, Roy. "The organization of decentralized information processing." *Econometrica* 61(5): 1109–1146, 1993.

[26] Rubinstein, Ariel. *Economics and Language*. Cambridge University Press, 2000.

[27] Claude E. Shannon. "A mathematical theory of communication." *Bell System Technology Journal* 27: 379–423, 1948.

[28] Simester, Duncan and Knez, Marc. "Direct and Indirect Bargaining Costs and the Scope of the Firm." *Journal of Business*, 75(2): 283-304, 2002.

[29] 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.

[30] Stein, Jeremy. "Information Production and Capital Allocation: Decentralized vs. Hierarchical Firms" *Journal of Finance*, 62(5): 1891-1921, 2002.

[31] 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.

[32] Wernerfelt, Birger. "Organizational Languages", *Journal of Economics and Management Strategy*, 13 (3): 461-72, 2004.

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

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

which proves the result. □

Proposition 2. *If the function d is "convex", i.e., if*

$$d(n+1) - d(n) \geq d(n'+1) - d(n') \text{ for all } n \geq n' \geq 1, \tag{A.1}$$

then, 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

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

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

Proposition 3. *If f represents a more complex environment than $\tilde{f}$, the minimal diagnosis cost associated with f is (weakly) larger than the minimal diagnosis cost associated with $\tilde{f}$, that is* $\min_{\mathcal{C}} D(\mathcal{C}; f) \geq \min_{\mathcal{C}} D(\mathcal{C}; \tilde{f})$.

Proof. By Proposition 1, given distribution f, there exists an optimal language which groups events into words only if they are adjacent according to indexing i. The optimal K-word language can be written as $(i_0, ..., i_K)$: the word k comprises all events with index between $i_{k-1} + 1$ and i_k (with the convention that $i_0 = 0$). The diagnosis cost induced by the optimal K-word language is:

$$\min_{\mathcal{C}} D(\mathcal{C}; f) = \sum_{k=1}^{K} \left(F_{i_k} - F_{i_{k-1}}\right) d\left(i_k - i_{k-1}\right).$$

By Proposition 1, it must be that $d\left(i_k - i_{k-1}\right)$ is nonincreasing in k. By re-arranging terms and noticing that $F_{i_0} = 0$ and $F_{i_K} = 1$, we can re-write the diagnosis cost as

$$\min_{\mathcal{C}} D(\mathcal{C}; f) = \sum_{k=1}^{K-1} F_{i_k} \left(d\left(i_k - i_{k-1}\right) - d\left(i_{k+1} - i_k\right)\right) + d\left(i_K - i_{K-1}\right)$$

But then we have

$$\begin{aligned}
\min_{\mathcal{C}} D(\mathcal{C}; \tilde{f}) &\leq \sum_{k=1}^{K} \left(\tilde{F}_{i_k} - \tilde{F}_{i_{k-1}}\right) d\left(i_k - i_{k-1}\right) \\
&= \sum_{k=1}^{K-1} \tilde{F}_{i_k} \left(d\left(i_k - i_{k-1}\right) - d\left(i_{k+1} - i_k\right)\right) + d\left(i_K - i_{K-1}\right) \\
&\leq \sum_{k=1}^{K-1} F_{i_k} \left(d\left(i_k - i_{k-1}\right) - d\left(i_{k+1} - i_k\right)\right) + d\left(i_K - i_{K-1}\right) \\
&= \min_{\mathcal{C}} D(\mathcal{C}; f)
\end{aligned}$$

where the first inequality is due to the fact that the optimal language for f may not be optimal for $\tilde{f}$ and the second inequality is due to the fact that f is more complex than $\tilde{f}$ and that $d\left(i_k - i_{k-1}\right) - d\left(i_{k+1} - i_k\right)$ is nonnegative for every k. □

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 $C_A = C_B$.*

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,[31] and, without loss of generality, assume that it belongs to $\mathcal{C}_A$.

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. □

Lemma 1. *For any two distributions f_A and f_B and any $\alpha \in [0,1]$, we have*

$$D^*(\alpha f_A + (1-\alpha) f_B) \geq \alpha D^*(f_A) + (1-\alpha) D^*(f_B).$$

Proof. Let us call $\mathcal{C}^*(f)$ the optimal code associated with distribution of events f, we have

$$\begin{aligned} D^*(\alpha f_A + (1-\alpha) f_B) &= \min_{\mathcal{C}} D(\alpha f_A + (1-\alpha) f_B, \mathcal{C}) \\ &= D(\alpha f_A + (1-\alpha) f_B, \mathcal{C}_*(\alpha f_A + (1-\alpha) f_B)) \\ &= \alpha D(f_A, \mathcal{C}^*(\alpha f_A + (1-\alpha) f_B)) \\ &\qquad + (1-\alpha) D(f_B, \mathcal{C}^*(\alpha f_A + (1-\alpha f_B)) \\ &\geq \alpha D^*(f_A) + (1-\alpha) D^*(f_B). \end{aligned}$$

□

Proposition 6. *An integrated form becomes relatively more profitable compared to the segregated form when either a) the diagnosis cost λ decreases, b) the synergy q^C/q^{NC} between the two services increases, or c) the heterogeneity of the two client distributions decreases.*

[31] That is $k \in \operatorname{argmin}_{\tilde{k}} n_{\tilde{k}}$ subject to $W_{\tilde{k}} \in \mathcal{C}_1 \cup \mathcal{C}_2$ and $W_{\tilde{k}} \notin \mathcal{C}_1 \cap \mathcal{C}_2$.

Proof. To prove this result, we use first the result in Lemma 1 to compare the effects of merging two different client populations. When homogeneity increases, 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.

We then require the following intermediate result: □

Lemma A.1. *If the client distribution* $(\tilde{f}_A, \tilde{f}_B, \tilde{\nu}_A, \tilde{\nu}_B)$ *is more homogeneous than* (f_A, f_B, ν_A, ν_B)*, then* $\tilde{\nu}_A D^*(\tilde{f}_A) + \tilde{\nu}_B D^*(\tilde{f}_B) \geq \nu_A D^*(f_A) + \nu_B D^*(f_B)$*. Consequently, the loss that results from a joint code is smaller in the more homogeneous client distribution.*

(Note that in both cases, the code of the integrated service has diagnosis cost $D^*(f)$; thus the second statement follows directly from the first.)

Proof. By definition 2, we have (note that we suppress, as in the text x from the notation f_x for simplicity)

$$\tilde{\nu}_A \tilde{f}_A + \tilde{\nu}_B \tilde{f}_B = \nu_A f_A + \nu_B f_B \text{ for all } x; \tag{A.2}$$

and

$$\begin{aligned} \tilde{f}_A &= \alpha_A f_A + (1-\alpha_A) f_B \text{ for some } \alpha_A \in (0,1) \text{ and all } x, \\ \tilde{f}_B &= (1-\alpha_B) f_A + \alpha_B f_B \text{ for some } \alpha_B \in (0,1) \text{ and all } x. \end{aligned} \tag{A.3}$$

Substituting (A.3) in (A.2), we obtain, for all x,

$$(\tilde{\nu}_A \alpha_A + \tilde{\nu}_B (1-\alpha_B)) f_A + (\tilde{\nu}_A (1-\alpha_A) + \tilde{\nu}_B \alpha_B) f_B = \nu_A f_A + \nu_B f_B$$

Because $f_A \neq f_B$, this implies

$$\begin{aligned} \tilde{\nu}_A \alpha_A + \tilde{\nu}_B (1-\alpha_B) &= \nu_A, \\ \tilde{\nu}_A (1-\alpha_A) + \tilde{\nu}_B \alpha_B &= \nu_B. \end{aligned}$$

We have

$$\begin{aligned} \tilde{\nu}_A D^*(\tilde{f}_A) + \tilde{\nu}_B D^*(\tilde{f}_B) &\geq \tilde{\nu}_A [\alpha_A D^*(f+A) + (1-\alpha_A) D^*(f_B)] \\ &\quad + \tilde{\nu}_B [(1-\alpha_B) D^*(f_A) + \alpha_B) D^*(f_B)] \\ &= [\tilde{\nu}_A \alpha_A + \nu'_B (1-\alpha_B)] D^*(f_A) \\ &\quad + [\tilde{\nu}_A (1-\alpha_A) + \tilde{\nu}_B \alpha_B] D^*(f_B) \\ &= \nu_A D^*(f_A) + \nu_B D^*(f_B), \end{aligned}$$

which proves the result. □

This result almost immediately yields (c) in the proposition. The other parts are also immediate.

Proof. We have

$$\frac{\Pi^I}{\Pi^S} = \frac{q^C(1-\lambda D^*(\nu_A f_A + \nu_B f_B))}{q^{NC}(1-\lambda(\nu_A D^*(f_A) + \nu_B D_*(f^B)))}.$$

Then, part a) of the proposition is an easy consequence of lemma 1, and part b) is obvious by examination.

As the client distribution becomes less heterogenous the profits under integration stay constant by (A.2) whereas the profits under separation increase by lemma A.1. This proves part c). □

Proposition 7. *For all $\mu \in (0, \mu^*)$, there exists $\lambda^{IH}(\mu)$ and $\lambda^{HS}(\mu)$ such that the unique optimal organization is*

$$\begin{array}{ll} \textit{integrated} & \textit{if } \lambda < \lambda^{IH}(\mu), \\ \textit{hierarchical} & \textit{if } \lambda \in \left(\lambda^{IH}(\mu), \lambda^{HS}(\mu)\right), \\ \textit{separated} & \textit{if } \lambda > \lambda^{HS}(\mu). \end{array}$$

The function λ^{IH} is decreasing, while λ^{HS} is increasing.

Proof. Let us call D^I, D^S, and D^H the values of D^* corresponding to the three organizational structures; by lemma 1

$$D^S < D^H < D^I. \tag{A.4}$$

Writing explicitly the dependance of profits on the parameters λ and μ, the profits corresponding to the three organizational structures are

$$\begin{aligned} \Pi^I(\lambda) &= q^C\left(1-\lambda D^I\right), \\ \Pi^S(\lambda) &= q^{NC}\left(1-\lambda D^S\right), \\ \Pi^H(\lambda,\mu) &= q^C\left(1-\lambda D^H\right) - \mu. \end{aligned}$$

The profits under the integrated and the separated organizations are equal for $\lambda = \lambda^{IS}$, where λ^{IS} is given by

$$\lambda^{IS} = \frac{q^C - q^{NC}}{q^C D^I - q^{NC} D^S}.$$

Equation (A.4) and $q^C > q^{NC}$ imply $0 < \lambda^{IS} < 1/D^I$. Equation (A.4) also implies $\Pi^I(\lambda) < \Pi^H(\lambda, 0)$ for all $\lambda < \lambda^{IS}$. Furthermore, for

$$\mu^* = \frac{q^C\left(q^C - q^{NC}\right)\left(D^H - D^I\right)}{q^C D^I - q^{NC} D^S} > 0,$$

we have

$$\Pi^I(\lambda^{IS}) = \Pi^S(\lambda^{IS}) = \Pi^H(\lambda^{IS}, \mu^*),$$

and therefore, for $\mu \in (0, \mu^*)$,

$$\Pi^I(\lambda^{IS}) = \Pi^S(\lambda^{IS}) < \Pi^H(\lambda^{IS}, \mu).$$

Because the slope of Π^H considered as a function of λ is intermediate between the slopes of Π^I and Π^S ($q^{NC}D^S < q^C D^H < q^C D^I$), the theorem is proved with

$$\lambda^{HS}(\mu) = \frac{q^C - \mu - q^{NC}}{q^C D^H - q^{NC} D^S}$$
$$\lambda^{IH}(\mu) = \frac{\mu}{q^C \left(D^I - D^H\right)}.$$

□

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[32] $\{[t_{k-1}, t_k]\}_{k=1,K}$.

The best K-words code is solution of

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

subject to

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

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 $k = 1, \ldots, K$.*

[32] 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 λ 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 know 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

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

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. □