The Cultural Shaping of ICTs within Academic Fields: Corpus-based Linguistics as a Case Study

Jenny Fry
Networked Research and Digital Information, NIWI-KNAW
Joan Muyskenweg 25, P.O. Box 95110, 1090 HC Amsterdam, The Netherlands.
jenny.fry@niwi.knaw.nl

Forthcoming in Literary and Linguistic Computing (Published by Oxford University Press on behalf of the Association for Literary & Linguistic Computing).

The aim of this paper is to show that the appropriation of ICTs is determined by a field’s specific cultural identity. Knowledge is not a homogenous whole, but a patchwork of heterogeneous fields. These fields are most visible as embodied in academic disciplines, which have distinct cultural identities shaped by intellectual and social considerations. Scholarly communication systems evolve over time within the context of these cultural identities. The paper discusses the cultural shaping of ICTs by drawing on an ongoing ethnographic study within corpus-based linguistics. The findings suggest that cultural elements such as ‘task-uncertainty’, ‘mutual-dependency’, heterogeneity, and institutional configurations will influence the appropriateness of a specific ICT infrastructure for a particular intellectual community. For example, fields that have a highly politicised and tightly controlled research culture will develop a coherent field-based strategy for the uptake and use of ICTs, whereas domains that are pluralistic and have a loosely organised research culture will appropriate ICTs in an ad-hoc localised manner. These findings demonstrate that overlooking cultural diversity in the development and implementation of ICT infrastructures and policies could prove detrimental for fields that do not map onto ‘big science’ conceptualisations of knowledge production. Furthermore, the paper demonstrates that effective understandings similarity and difference in patterns of scholarly communication need to take the fine-grain of specialist fields as the unit of analysis, rather than the course-grain of the discipline.
Communication is central to the academic enterprise... [it] is the force that binds together the sociological and epistemological, giving shape and substance to the links between knowledge forms and knowledge communities. (Becher, 1989, p. 77)

1 Introduction

The past two decades have been host to an explosion in information communication technologies (ICTs). This has created a cornucopia of digital networks and resources connected on a global scale. Scholars are no longer limited to the annual meetings of scholarly associations and societies to communicate informally with their national and international peers. The opportunity is there for them to stay in touch with their colleagues and fields through a plenitude of email networks. Availability of channels for the formal communication of scholarly work has expanded far beyond the local collections of academic libraries. Scholars no longer have to travel to the information source, as the information source can be delivered digitally to their desktop. There has been a great deal of speculation about the impact of digital communication media, such as the Web, on the work practices of scholars. We are told that with the arrival of the Internet there has been an ‘information revolution’ that will potentially alter scholarly communication and knowledge production in radical ways. However, there is a need to develop a grounded understanding of how scholars are actually using ICTs in their work.

Studies investigating scholarly communication and the use of ICTs have tended to focus on the physical and applied sciences to the neglect of humanities fields. There are exceptions in the research literature to this preoccupation, including Stone (1982), Brockman et al (2001), Palmer and Neumann (2002), and (Talja, 2002) who have studied information seeking behaviours of humanities scholars in the digital library realm. This paper helps to address this lacuna in research into scholarly communication by drawing on earlier comparative research spanning the humanities, social sciences and physical sciences (Fry, 2003), and current research focusing particularly on specialist fields within linguistics. Development of broad reaching ICT policy initiatives, such as the e-science programme in the UK that has invested millions of pounds into the promotion of digital network projects in science, means that it is more important than ever to have an empirical understanding of how intellectual fields outside of the physical sciences use networked digital resources. Failure to develop ICT policy that is informed by systematic comparative research could seriously disadvantage fields that do not map onto the physical science model of communication and knowledge production. In response to increasing policy initiatives, such as e-science and GRID infrastructures, across Europe and North America the aim of the paper is to illustrate how the
cultural identity of intellectual fields influences ways in which ICTs are appropriated. Implicit in terms such as ‘e-science’, ‘e-humanities’ and the ‘semantic web’, and notions of ‘digital scholarship’ in general, is the collaborative imperative. Indeed, during the DRH 2003 conference the rhetoric of collaboration resonated throughout a number of the sessions, most strongly so in the panel discussion on ‘e-science’ chaired by Sheila Anderson. Though there were many strident declarations of the need for more collaborative work in the humanities few papers were explicitly concerned with understanding the process of distributed collaboration or the development of collaborative technologies around these processes. Therefore, a subsidiary aim of this paper is to demonstrate the need for a reflexive approach to the development and implementation of ICT infrastructures within intellectual fields. The paper addresses these issues by discussing theories on academic research cultures in the context of an ongoing ethnographic study of the field of corpus-based linguistics and the communities that inhabit it.

2 Current Understanding in the Disciplinary Shaping of Scholarly Communication

Current understanding in research into scholarly communication indicates that a range of social and cultural conditions will influence the uptake and use of computer-mediated communication technologies within scholarly communities. For example, Orlikowski and Gash (in Kling and Lamb, 1996, p.48) found that ‘people’s fine-grained work incentives influence whether they see technologies as relevant, and the ways in which they appropriate the technologies’. Kling, Spector and McKim (2002) illustrate how the cultural context of disciplines can lead to the rejection of digital resources. They observed that attempts made by the National Institutes of Health (NIH) in North America to implement a digital pre-print server model of publishing (arXiv.org) in bio-medical science, which is popular within the disciplines of physics, mathematics and chaos theory, were resisted by lead scientists in the field. They believe that resistance to the model can be attributed to disparity between the pre-existing model of publishing in bio-medical science and the culture of publishing inherent in the digital pre-print server model.

Comparison of influential cultural factors across scientific fields is problematic due to the multifaceted nature of scientific cultures. For example, differences in patterns of communication for collaboration have been accounted for from a variety of perspectives. Kling, Spector and McKim (2002)
used publishing models as a frame of reference, Olson and Olson (2000) were interested in the affects of geographic distance, and Kraut, Galegher and Egido (1988) examined the influence of interpersonal factors. All such perspectives are valid, but constrained in that they each only account for a limited number of facets of a research community’s complex cultural identity, so are limited in their predictive powers. Additionally, research in information science, the intellectual tradition from which I write, is limited in understanding the cultural construction of patterns of communication, because it tends to focus on communication products, such as journal articles, rather than on informal communication at the process level, such as collaborative practices.

2.1 Cultural identity of scholarly communities

Academic disciplines have distinct cultural identities, which are shaped by the social, ‘establishment of reputation’, and the epistemological, ‘promotion of knowledge’, elements of scholarly activity (Becher 1989). Becher claims that disciplines, e.g. physics, anthropology and history, can be conceived as having recognisable identities and particular cultural attributes. He also emphasises the close interrelationship that exists between the characteristics and structures of knowledge fields, and the attitudes, activities and cognitive styles of the communities of academics that occupy them. In his seminal work Becher (1989) studied twelve academic disciplines in order to investigate the relationship between the social and the intellectual elements of knowledge fields. After a series of interviews with two hundred and twenty-one academics he devised a taxonomy, which can be used to characterise social behaviours and knowledge structures across academia. The taxonomy is based upon earlier systems developed by Biglan (1973) and Kolb (1981) who were concerned with the extent to which knowledge structures can be classified as hard/soft and pure/applied. The four broad categories of the taxonomy are summarised in table 1 below.

<table>
<thead>
<tr>
<th>Group</th>
<th>Knowledge</th>
<th>Culture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Sciences</td>
<td>Cumulative; atomistic (crystalline/tree-like); concerned with universals,</td>
<td>Competitive, gregarious; politically well organised; high publication</td>
</tr>
<tr>
<td>(e.g. physics)</td>
<td>quantities, simplification; resulting in discovery/explanation</td>
<td>rate; task oriented</td>
</tr>
<tr>
<td>‘hard-pure’</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Jenny Fry-The cultural shaping of ICTs within academic fields: corpus-base linguistics as a case study-forthcoming in *Literary and Linguistic Computing*

| Humanities (e.g. history) & Pure Social Sciences (e.g. anthropology) | Reiterative; holistic (organic/river-like); concerned with particulars, qualities, complication; resulting in understanding/interpretation | Individualistic, pluralistic; loosely structured; low publication rate; person oriented |
| Applied Sciences (e.g. mechanical engineering) | Purposive, pragmatic (know-how via hard knowledge); concerned with mastery of physical environment; resulting in products and techniques | Entrepreneurial, cosmopolitan; dominated by professional values; patents substitutable for publications; role oriented |
| Applied Social Sciences (e.g. education) | Functional, utilitarian (know-how via soft knowledge); concerned with enhancement of [semi-] professional practice; resulting in protocols and procedures | Outward looking; uncertain in status; dominated by intellectual fashions; publication rates reduced by consultancies; power oriented |

**Table 1 Matrix of disciplinary groupings (adapted from Becher, 1987, p.289)**

Becher’s typology is a useful starting point to think about points of divergence for disciplinary culture, but it has its limitations. A major limitation is its basis on the discipline as a unit of analysis, which makes it an interesting descriptive tool, but too crude to be used as an analytic device. Corpus-based linguistics illustrates how the typology glosses over the complexity of cultural types at the specialist field level and certainly does not account for interdisciplinary fields that fall between the interstices of the physical, applied, social sciences and humanities. Corpus-based linguistics cannot be mapped into any one category in the typology in a mutually exclusive way, rather it straddles the boundary between the pure social sciences and applied sciences. With its mix of theoretical ‘soft-pure’ influences and its technical ‘hard-applied’ influences, it is both holistic and pragmatic, concerned with interpretation of linguistic phenomena and the development of products and techniques.

Turning to the scholarly communities that inhabit disciplines, Becher draws an analogy between ‘urban’ and ‘rural’ ways of life to describe major ‘lifestyle’ differences. According to this analogy, communities that have a high people-to-problem ratio, such as high-energy physics, are likened to urban populations and those that have a low people-to-problem ratio, such as modern linguistics, are likened to rural populations:

In the first, there is alongside a densely concentrated population a generally busy – occasionally frenetic – pace of life, a high level of collective activity, close competition for space and resources, and a rapid and heavily used information network. By and large, the rural scene, though it may offer frenetic and competitive moments, occasions for communal involvement and a potential for spreading rumour and gossip like wildfire, displays the opposite characteristics. (Becher, 1989, p.79)
Becher suggests that inhabitants of ‘urban’ specialisms work within narrowly constrained areas of intellectual enquiry, while those in ‘rural’ specialisms engage in a broader range of problems within a less well-defined intellectual territory. Within ‘urban’ areas large numbers of researchers cluster around a small number of high-profile topics, while in contrast there is a wide dispersion of topics and researchers in ‘rural’ areas. ‘Urban’ communities are concerned with producing results with alacrity, whereas ‘rural’ communities are concerned with research questions that require a more ponderous approach. Competition is a prominent characteristic of ‘urban’ life, while ‘rural’ communities are more likely to opt for a division of labour. Due to the high people-to-problem ratio in ‘urban’ communities collaboration is more likely to occur than in ‘rural’ areas, where there is less overlap between people and problems. According to Becher, the research environments of ‘urban’ and ‘rural’ areas differ just as greatly as the activities that occur within them. Becher argues that the distinction between behaviours amongst the communities that inhabit ‘urban’ and ‘rural’ areas emerge most clearly in their modes of communication. This is illustrated by the higher premium placed on informal ‘word of mouth’ communication in ‘urban’ areas, such as the prevalence of pre-print systems, and the relatively infrequent attendance at international conferences by specialists in ‘rural areas’. Here, again, we see a crudeness in the level of granularity adopted by Becher. To a certain extent these limitations are also related to the physical science centric orientation in which theories of scholarly communication are developed, which means that fields outside of the physical science tradition are underdetermined. Theory tends to be based around characteristics of ‘big science’, with fields outside being treated homogenously in their absence of physical science characteristics. For example, the limited number of ‘big science’ disciplines are ‘hard pure’, created by convergent and urban communities, and all other disciplines and communities are ‘soft’ or ‘applied’, divergent and rural, but what about urban research styles in soft/applied disciplines, such as Health Psychology?

2.2 An explanatory framework

The most powerful explanatory device for putting disciplines under the microscope is the theory developed by Whitley (1984). Considering both epistemological and social structures within disciplines he explains cultural difference across science based on two dimensions of scholarly practices, ‘mutual dependency’ and ‘task uncertainty’. ‘Mutual dependency’ relates specifically to the extent to which a field is dependent upon
knowledge produced in other fields in order to make a significant contribution to science. It also includes the degree of interdependence between scientists and groups in a field. For example, the extent to which scientists’ are dependent upon particular groups of colleagues to make competent contributions to collective intellectual goals and acquire prestigious reputations that lead to material rewards, such as grants. Additionally, it accounts for the extent to which a field adopts evaluation criteria and standards from other fields for the assessment of work produced outside its intellectual boundary, rather than developing its own criteria for evaluating externally produced knowledge. ‘Task uncertainty’, on the other hand, concerns the unpredictability of task outcomes. Whitley argues that because the sciences are committed at an institutional level to produce novel results, research activities are ‘uncertain compared to other work activities’ in that ‘outcomes are not repetitious and highly predictable’. Whitley uses 20th Century chemistry as an example of a field that has higher levels of ‘mutual dependence’, but lower levels of ‘task uncertainty’, while he uses sociology as an example of lower levels of ‘mutual dependence’, but higher levels of ‘task uncertainty’. Furthermore, Whitley stratifies ‘mutual dependence’ into functional and strategic factors and ‘task-uncertainty’ into technical and strategic factors. Functional and Technical facets tend to relate to the research object and strategic facets tend to be determined by social, institutional and political considerations.

A limitation of the framework is that levels of ‘mutual dependence’ or ‘task uncertainty’ cannot be measured, rather it is used to position fields intellectually and socially in relation to one another, which does however make it a useful device for comparative research. Though Whitely (1984) also developed his theory of cultural identities across science based on the discipline as the unit of analysis the axes of ‘mutual dependency’ and ‘task uncertainty’ address complexity around epistemological and social concerns in more detail than Becher’s axes of ‘hard/pure’ and ‘soft/applied’, which are very much constrained to painting a broad brush across the sciences at the general level. Whitley’s theory better allows for more fine-grained analysis, because of his emphasis on degrees of either ‘task uncertainty’ or ‘mutual dependency’, which can be applied at a number of levels of granularity, including the specialist field. However, it should be noted that although Whitley does draw on the applied sciences, social sciences and humanities in developing his theory it is based on a limited number of disciplines. Coming from a business management perspective the disciplines that Whitely chose to base his theory upon tend to be older macho fields, such as
physics, chemistry and philosophy. Furthermore, except for business studies his examples of contemporary domains are conspicuous by their absence. None-the-less, this paper argues that Whitley’s theory can be used as an analytical framework to understand the cultural shaping of ICTs within intellectual fields. To illustrate how the theory can be applied to the study of scholarly communication and predict patterns of ICT appropriation it is discussed below in the context of corpus-based linguistics.

3 The Intellectual and Social Orientation of Corpus-based Linguistics

The purpose of this section is to describe the cultural identity of corpus-based linguistics as constructed by the community of linguists that inhabits it. The empirical data discussed below is based on case study research, which is a method involving the gathering of information about a particular person, social setting, event, or group, which enables the researcher to reach a detailed understanding of how the phenomenon being studied functions. The case of corpus-based linguistics was part of a series of case studies, including high-energy physics and social/cultural geography (Fry, 2003), and latterly the field of argumentation theory, so the findings are based on comparative research. The advantage of comparative research is that it can reveal insights that otherwise would not have light shed upon them. Consequently, the text occasionally makes reference to the other cases in order to throw the patterns of behaviour that I observed within corpus-based linguistics into relief.

The linguists upon whose accounts this chapter is based were intellectually located in a range of fields, such as computational linguistics, natural language processing (NLP), translation studies and contrastive linguistics, but they all identified themselves as being research active within corpus-based linguistics. All of the linguists were UK-based from six universities and eight departments across England and are referred to in the text as participants.

3.1 The research object

The community that inhabits corpus-based linguistics is concerned with the empirical analysis of naturally occurring language. It takes an empirical approach to language description using a number of analytical phenomena, such as discourse, morphology and phonology, applied to a wide variety of languages, as Lawler and Dry (1998) explain:
The methods that underlie corpus research are based upon distributional analysis and the study of co-occurrence data, decomposition of analysis into multiple layers of phonology, morphology, syntax and discourse, and automatable discovery of linguistic descriptions or grammars (Lawler and Dry, 1998).

Empirical evidence from the perspective of corpus linguistics is naturally occurring instances of language use. In order to study this data, corpus-based linguists build and use large databases containing examples of spoken or written language, such as the British National Corpus (BNC), the Corpus of Spoken Dutch (CGN), and the International Corpus of English (ICE). Tasks typically involved in building a corpus include corpus design, recording, transcription, tagging, annotation and the development of exploitation software. There are many decisions to be made about the appropriate approach to each of these tasks. For example, what level the annotation should be at, syntactic or prosodic? In fact, the process of corpus annotation is a major element of knowledge production in the field and is reported by corpus-based linguists in general as extremely problematic. They report a high degree of ‘technical task uncertainty’, with annotation relying heavily upon the subjective judgement of the annotator based on her tacit knowledge of the language under study. Atwell et al, (2002) conducted a survey of the use of parsing standards within the field and found that there was much local variation in the use of parsing schemes:

The rather disheartening conclusion we can draw from these observations is that it is difficult, if not impossible to map between all the [parsing] schemes… No single standard can be applied to all parsing projects. Even the presumed lowest common denominator, bracketing, is rejected by some corpus linguists… The guiding factor in what is included in a parsing scheme appears to be the author’s theoretical persuasion or the application they have in mind. (Atwell et al, 2002)

In order to overcome issues of subjectivity and produce empirically based ‘facts’ about linguistic structures in a text it is common practice for a number of analysts to mark up a text independently and use statistical formulae to measure how much they agree. The idea is that subjective bias can be eliminated by only admitting annotations where there is a high agreement score. This typically takes several iterations, re-defining the task each time until a satisfactory level of agreement is reached.

Whitley observes that in fields where there is a high degree of ‘technical task uncertainty’ results will be ambiguous and subject to a variety of conflicting interpretations and the use of technical procedures will be highly tacit, personal and fluid. Whitley argues that the implications of higher levels of ‘technical task uncertainty’ for the organisation and control of research are an increased reliance upon direct and personal control of how research is conducted, local variations in work goals and processes, and greater
emphasis upon informal communication and coordination processes. We can observe this characterisation played in patterns of collaborative work organisation in corpus-based linguistics (see Section 3.4).

3.2 Disciplinary boundaries

When corpus-based linguistics was in its infancy, during the 1970s and 1980s, it was considered to be on the margins of the linguistics discipline, but during the past decade its approaches and perspectives have become more widely accepted within mainstream linguistics. This acceptance has, no doubt, been assisted by developments in computing capabilities and an increasing emphasis upon evidence-based research within linguistics. Even those areas that overlap with corpus-based linguistics have traditionally been viewed as peripheral to linguistics. NLP for example, was perceived by the disciplinary community to benefit from linguistic theory and method, while making a limited contribution to linguistics itself. Then in the mid 1980s when NLP techniques and products became practicable and useful to ‘real world’ users, its relationship with linguistics shifted. Lawler and Dry (1998) argue that the construction of robust NLP systems exposed linguistic theory and method to testing and validation, which lead to its prominence in the discipline. Consequently, they observed an increasing trend within the discipline to prove that ‘linguistic theories are computationally effective’, thus increasing the degree of strategic dependence.

Inhabiting an interdisciplinary space, the linguists had some difficulty identifying themselves with a particular field designation. There was variation amongst the linguists in terms of how they labelled their intellectual territory. This contrasts with a monodisciplinary field such as high-energy physics, whereby, its inhabitants consistently identified themselves as high-energy physicists. Linguists working in the same research area, on the other hand, may choose to identify themselves with different designations, while those working in different areas may choose to identify themselves with the same designation. For example, one participant from NLP explained that it was not uncommon for him to have very little in common with researchers calling themselves corpus-based linguists, because they may be studying literary styles or English language learners, but can have a lot in common with researchers who would not necessarily identify themselves as corpus-based linguists.

One of the computational linguists argued that the way in which corpus-based linguists choose to identify themselves is contingent upon the nature of the audience they are addressing and wish to align
themselves with. He explained that because his background was not actually in computational linguistics he would choose to identify himself as a computationally oriented linguist, rather than a computational linguist within computational linguistics circles.

Identifying with a particular field designation was further complicated by the fact that many of the linguists were pursuing more than one research direction at a time. During their academic careers it was typical for the linguists to move from one field to another. Being academic ‘nomads’, many of the corpus-based linguists had travelled, intellectually speaking, a long way from their original roots in linguistics. This was reflected in the fact that almost all the linguists were situated within academic departments that had a technical, rather than a linguistic orientation, such as language engineering. Consequently, whilst many of them were in departments where they could obtain resources for the technical aspects of their research, there was a lack of resources for the theoretical aspects. The opposite situation was true for the few corpus-based linguists who were in more traditional departments of linguistics, such as phonetics. A number of the linguists reported that connections developed in mainstream linguistics earlier in their careers had ceased to exist since moving to technical departments.

Interdisciplinarity and cross-border movement means that the intellectual boundaries of corpus-based linguistics are ‘fuzzy’ and that there are multiple audiences for research output. Whitley associates these characteristics with low levels of ‘strategic mutual dependency’. He observes that in fields where the degree of ‘strategic mutual dependence’ is relatively low researchers are able to make contributions to a variety of goals without needing to incorporate specific results and ideas of particular specialist colleagues in a systematic way. Researchers tend to pursue a variety of intellectual objectives with a variety of technical approaches and so contributions can be quite diverse.

### 3.3 Institutional structures

With the exception of seminar series the opportunities for intellectual social interaction within individual universities were rather limited. Invited speakers and other academic visitors tended to add much needed critical mass to the intellectual activities of departments. It is clear that concentrations of critical mass play an important role in sustaining scholarly communities through interchange of knowledge and expertise.
Nationally, there is a lack of critical mass within corpus-based linguistics and this is reflected in the lack of institutional and organisational representation of the field in the UK, such as academic departments, schools, professional associations and journals. Over the last decade several attempts have been made to establish networks of researchers across the UK and Europe-wide in domains that closely relate to corpus-based linguistics, but not directly in the field itself. For example, there is SALT (Speech and Language Technology club in the UK), CLUK (The UK special-interest group for computational linguistics) and ELSNET (European Network of Excellence in Human Language Technologies). According to the CLUK manifesto\(^1\) the focus of SALT has been on joint speech and language work, and on collaborative industrial and academic work, while these are important to the NLP community it does not provide a platform for issues pertaining specifically to NLP. Therefore, CLUK emerged in response to what the NLP community considered as a need for EPSRC (Engineering and Physical Science Research Council), a major funding body in the UK, to hear the voice of the community.

ELSNET\(^2\) was established in 1991 as part of the European Commission Networks of Excellence scheme, again, to mainly support the language and speech technology community. The ELSNET mission statement says that ‘It operates in an international context, and will consider, across discipline boundaries, all human communication research areas related to language and speech’ (Extract from the ELSNET mission statement, September 2001).

The creation of an association such as CLUK demonstrates a need to create fora for sustaining and giving voice to the corpus-based linguistics community. These institutional configurations mean that the audience for the field is quite diffuse, which complicates the development of significance standards and has significant implications for the allocation of resources. We see this where feelings of anxiety were expressed with regard to the field’s interdisciplinary nature and the Research Assessment Exercise in the UK (see Section 5). These observations relate to Whitley’s theory of ‘strategic mutual dependence’, which is the extent to which researchers have to persuade colleagues of the significance and importance of their problem and approach to obtain a high reputation from them. A consequence of higher levels of ‘strategic mutual dependence’ is that coordination goes beyond the technical level of integrating specialist

---

\(^1\) CLUK manifesto available at, http://www.dcs.shef.ac.uk/research/cluk/manifesto.html

\(^2\) Details of ELSNET available on the website at, http://www.elsnet.org/
contributions to a common goal, rather it is a political activity that involves the organisation of programmes and projects in terms of particular priorities and interests. The political imperative culminates in the setting of research agendas, allocation of resources, and the influence over careers.

3.4 Collaboration and coordination

Although reports of collaborative work amongst the case study participants were plentiful, so also were reports of individual intellectual endeavours, with some linguists stating that it was more usual for them to work independently. Some of the linguists described themselves as lone independent researchers and this is most likely a function of their theoretically oriented research interests, which tend to be polemical and individualistic in nature.

Collaborative groupings in corpus-based linguistics range from the institutionally bounded to the trans-national. The two main types of formal collaborative activity that the linguists reported participating in were the development of corpora and the production of publications. A number of challenges to collaborative work are apparent within the field. One the major challenges being the sheer scale of the technical resources required in building large national and international corpora. Co-ordination was a particular challenge for the widely distributed European and international collaborations concerned with building multi-lingual parallel corpora. Patterns of work organisation for collaboration were very localised. The corpus-based linguists reported that informal face-to-face communication was essential for coordinating European and international research projects, and that due to the wide-distribution of participants such communication was very limited.

Several ‘functional mutual dependencies’ exist between the fields that overlap within corpus linguistics, such as linguistics and computer science, statistics and linguistics, natural language processing and linguistics. For example, in order for an end user to use a parallel corpus it is necessary that the text material in the different languages is aligned computationally. Alignment is the process by which the same text in different languages are linked to each other, e.g. at the sentence level, so that it is easy to find the parallel text in the other language(s). This is one aspect that relies on substantial input from the NLP community. Another aspect of interdependency in the domain is related to the multi-lingual element of many of the corpora being developed. Researchers involved in building international multi-lingual corpora
will typically be knowledgeable in one or two languages only. Thus they will be dependent on research partners to contribute national corpora to international collaborations. These large-scale collaborations are usually based upon clear divisions of labour according to the natural languages of each partner’s region. The participants reported that co-ordination of these widely distributed collaborations was extremely problematic. Difficulties can occur with the flow of information and face-to-face meetings played a crucial role in overcoming these issues and coordinating groups. For example, one participant reported problems related to the communication of copyright information regarding the publications that were being used in the corpus across the various participating countries. The linguists also reported difficulties in information flow arising from the variety of computer platforms and software being used by partners in large collaborations. This was compounded by variations in the national ICT infrastructures being used, with some partner countries having very limited, or no access at all, to the internet, and consequent lags in communication had meant that the partners often fell behind in the development schedule.

There seems to be some doubt expressed by the participants with regard to the relative success of these widely distributed projects. According to some, while the projects they had been involved in had produced useful research they had not succeeded in delivering a working system within the allocated funding period. This meant that a number of participants were struggling to continue the research without funding.

The key processes of formal collaboration such as planning, decision-making, and the production of artefacts are all dependent on face-to-face conversations. Often ideas are not crystallised and moved forward until collaborative partners are able to meet directly. Most of the participants reported that they resolved this issue by using annual conferences, such as ICAME (International Computer Archive of Modern and Medieval English), as sites of convergence for collaborative partners. Physical proximity is an essential element of scholarly life on both a personal and collective level. It enables scholars to become acquainted with one another and build intellectual communities, providing essential support for those functions integral to their existence, such as collaboration, problem solving, and peer review.

In this section we have observed that corpus-based linguistics has a relatively high degree of ‘functional dependence’, but a lower degree of ‘strategic dependence’, coupled with a relatively high
degree of both ‘strategic’ and ‘technical task uncertainty’. The following section discusses the consequence of this cultural configuration for the uptake and use of ICTs within corpus-based linguistics.

4 Technology Use: The Challenge of Developing Standards

The lack of standardisation around the research object in corpus-based linguistics gets played out in the appropriation of ICTs. Although the corpus-based linguists recognized the importance of technical standards around their objects of research they had not succeeded in developing a unified system of technical and social standards and protocols for computer-mediated communication across the field. Use of the web and other computer-mediated communication technologies appears to be determined at the level of individual research groups, rather than on a community-wide basis, as is the case with high-energy physics. However, the attitude towards communication media reported amongst the corpus-based linguists was one of seeking improvements to existing technology and finding solutions to communication needs based upon perceived intellectual and social requirements. The linguists reported a number of ways in which computer-mediated communication had been used to assist inter-personal communication and collaboration. A variety of social and technical standards and protocols have evolved in the domain, however, consensus on which should be predominant had not been reached. In part, this can probably be attributed to the fact that corpus-based linguistics is influenced by quite distinct and different research cultures. For example, while LaTex is used in Lexicography and NLP, it is not a standard typically used within corpus-based linguists, which has implications for the downloading of papers from the web. Disparities can also exist in the availability of computer-mediated communication technologies across collaborators.

In the technical areas of corpus-based linguistics many resources are available digitally via the web. Even prior to the development of the web, resources such as bibliographic databases were available electronically and shared amongst the community. Traditionally, these resources were produced and maintained locally within individual academic departments and the culture of producing and organising digital resources locally has been transferred to the web. The popularity of departmentally produced technical report series or working papers series, which typically are publicly available on the web, are a prime example of locally produced shared digital resources in the field. As are the detailed and frequently cited web based annotated lists of online resources in the field. The tradition of local work organisation and
locally produced digital resources contrasts with the centralised approach adopted by the high-energy physics community.

We can observe a contrast in ICT use between those communities in the field that are technically oriented and those that are theoretically oriented. There has been a corresponding difference in the development of social and technical standards and protocols for computer-mediated communication across the field. The reason for such variation is that being an interdisciplinary domain corpus-based linguistics is constituted from a number of cultural types with differing social norms and intellectual concerns. Crossing both ‘hard-applied’ and ‘soft-pure’ disciplinary boundaries the social organisation of the field has both tightly-structured and loosely structured elements within it, leading to variation in levels of cultural identity, hence two distinct patterns of computer-mediated communication.

The corpus-based linguists were the only group amongst the other case studies to articulate explicitly a desire to promote their research and make it visible to the scholarly community through the web. The participants reported using three major channels for promoting their research, discussion lists, personal web pages and conferences. With regard to dissemination of research and information seeking behaviour, again, there is a differentiation between the theoretical and technical areas of the field. Practitioners within the domain of computational linguistics were more likely to make their papers available via personal web sites and expect others to make their research available in the same way. A commonly reported disincentive for making papers available in electronic format is the additional workload incurred. Research material is less likely to be available electronically in the more theoretical areas of the domain. When it comes to finding out about the state-of-the-art, conferences were cited as the most important forum for communication. In addition to being at the loci of cutting edge research, conferences enable the linguists to locate themselves within the intellectual territory of the community at large. In terms of published channels of communication conference proceedings were reported as the most important source of information for the technically oriented communities in the field, whereas, in the more theoretically oriented areas of the field peer-reviewed journal papers were considered to be more important. There was a perception amongst the linguists that currency and speed of publication were less of a priority in the theoretically oriented areas of the domain. Across the board discussion lists were a popular mechanism for keeping in touch with the wider community on a daily basis.
I observed that two pivotal points link Whitley’s (1984) theory of levels of ‘functional mutual dependence’ with the appropriation of ICTs within scholarly communities: coordination and standards. He argues that the degree of ‘functional dependence’ between members of a field determines the need to coordinate task outcomes and demonstrate adherence to common competence standards. Having established that high-energy physics is characterised by high levels of ‘functional dependence’ we can link Whitley’s prediction of how this is manifested to the highly coordinated way in which the high-energy physicists have developed field-based social and technical standards and protocols for ICTs and distributed collaborative work. Corpus-based linguists deal with a wide range of tasks and languages, and collaborations involve the development of large complex technical systems across a range of national research and technological infrastructures. This tends to result in uncertain task outcomes and a lack of standardised procedures. The corpus-based linguists reported that informal face-to-face communication was essential for coordinating European and international research projects, and that due to the wide-distribution of participants such communication was very limited. I argue that the lack of success across the corpus-based linguistics community in developing field-wide social and technical standards and protocols for computer-mediated collaborative work can be attributed to low levels of ‘strategic dependence’ and high levels of both ‘strategic’ and ‘technical uncertainty’.

Whitley does not discuss his theory in the context of ICTs, therefore, this paper extends his theory by arguing that similarities and differences in the ways in which intellectual fields adopt and develop ICTs can indeed be explained in the context of Whitley’s theory.

5 Cultural Factors not Addressed by Whitley

Whitley’s theory does not explain the whole story in the account given above. There are other influential cultural factors that fall outside of the scope of his taxonomy, such as levels of interpersonal recognition amongst community members, size of field, and institutional frameworks for the organisation of knowledge producing communities. Institutional structures can influence patterns of scholarly communication in a number of ways. This is because they have a potential affect on factors such as the formation of critical mass, recognition and reward, and quality control systems. There are also internal institutional influences at play that will determine patterns of communication and collaboration across a domain, such as the extent to
which a domain has become institutionalised. It is understood here that institutionalisation of a domain means the establishment of an extensive social network of researchers, the establishment of informal and formal fora for the communication and dissemination of ideas, representation within university and other organisational structures, and recognition at the international level. For example, Investigating the development of scientific specialities Mullins (1972) found that social structures form around phases of intellectual development. Using Kuhn’s notion of paradigm shifts, Mullins correlates the conceptual development of the Phage Group, which first appeared in the mid-1930s and developed into a specialty of molecular biology, with what Kuhn describes as phases of ‘paradigm development’, ‘success’, and ‘puzzle-solving’. Mullins classifies the corresponding social structures that he identified as ‘paradigm group’, ‘communication network’, ‘cluster’, and ‘specialty’. Mullins is keen to point out that the four social structures that he identified are not mutually exclusive and that there is considerable overlap in the social activities that take place within them:

Paradigm and network-type structures, in particular, continue to form and function even after a group has entered the cluster or specialty stage. Their function at these stages is usually to supply new members and ideas to the increasingly formal cluster or specialty, although occasionally a totally new cluster will result. The analytical distinction of the different stages is not intended to imply that the activities which constitute the preceding stages are no longer functioning; it is intended only to show that for a given intellectual problem, a more complex structure has been established. (Mullins, 1972, p.54)

Accepting Mullin’s theory that the intellectual development of a domain will have a specific corresponding social structure it is highly probable that the phase of intellectual development that a domain has reached will also influence the appropriateness of specific communication media and the nature of the resulting communication system.

With regard to recognition and reward many of the corpus-based linguists felt that the quality control system exercised by the RAE within the UK had a significant influence upon patterns of communication within the domain. Potential clashes of interest were reported by the linguists between research output measures, such as the RAE, and the interdisciplinary nature of the domain. One reason for this is the polarisation of communicative modes between those communities in the field that are technically oriented and those that are theoretically oriented, which means that kudos associated with certain forms of publication vary across the field.
6 Conclusions

When it came to operationalising Whitley’s and Becher’s axes of convergence and divergence as analytic devices they worked well for describing the patterns of ICT appropriation and use observed within the high-energy physics community, but were both insufficient in providing an illustrative explanation for the patterns I observed amongst the corpus-based linguists. A factor that compounds this limitation is that both typologies are based on the discipline as the unit of analysis, which accounts for why they are effective for describing a monodisciplinary specialism, such as high-energy physics, but less so for interdisciplinary specialisms such as corpus-based linguistics. Whitley is less limited in this sense than Becher, because the axes of ‘mutual dependence’ and ‘task uncertainty’ address complexity around epistemological and social concerns in more detail than Becher’s axes of ‘hard/pure’ and ‘soft/applied’, which are very much constrained to painting a broad brush across the sciences at the general level. Becher also presents types as more mutually exclusive than Whitley, who better allows for more fine-grained analysis, because of his emphasis on degrees of either ‘mutual dependence’ or ‘task uncertainty’. The advantage of this is that it can be applied at a number of levels of granularity, including the specialist field. Integrating Whitley’s theory of ‘mutual dependence’ and ‘task uncertainty’ with my empirical findings demonstrates that there is a connection between these two dimensions of scholarly culture and patterns of ICT appropriation within fields. This not only lends weight to current understanding that technology does not have an autonomous effect, rather it is appropriated differentially across intellectual fields, but also that we can use Whitley’s theory as a framework for further exploratory work in guiding the development and implementation of ICT infrastructures and information policy across diverse scholarly communities.

7 References


Table 1 Matrix of disciplinary groupings (adapted from Becher, 1987, p.289)

<table>
<thead>
<tr>
<th>Group</th>
<th>Knowledge</th>
<th>Culture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Sciences (e.g. physics)</td>
<td>Cumulative; atomistic (crystalline/tree-like); concerned with universals, quantities, simplification; resulting in discovery/explanation</td>
<td>Competitive, gregarious; politically well organised; high publication rate; task oriented</td>
</tr>
<tr>
<td>‘hard-pure’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Humanities (e.g. history) &amp; Pure Social Sciences (e.g. anthropology)</td>
<td>Reiterative; holistic (organic/river-like); concerned with particulars, qualities, complication; resulting in understanding/interpretation</td>
<td>Individualistic, pluralistic; loosely structured; low publication rate; person oriented</td>
</tr>
<tr>
<td>‘soft-pure’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applied Sciences (e.g. mechanical engineering)</td>
<td>Purposive, pragmatic (know-how via hard knowledge); concerned with mastery of physical environment; resulting in products and techniques</td>
<td>Entrepreneurial, cosmopolitan; dominated by professional values; patents substitutable for publications; role oriented</td>
</tr>
<tr>
<td>‘hard-applied’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applied Social Sciences (e.g. education)</td>
<td>Functional, utilitarian (know-how via soft knowledge); concerned with enhancement of [semi-] professional practice; resulting in protocols and procedures</td>
<td>Outward looking; uncertain in status; dominated by intellectual fashions; publication rates reduced by consultancies; power oriented</td>
</tr>
<tr>
<td>‘soft-applied’</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Jenny Fry-The cultural shaping of ICTs within academic fields: corpus-base linguistics as a case study-forthcoming in *Literary and Linguistic Computing*

1 CLUK manifesto available at, http://www.dcs.shef.ac.uk/research/cluk/manifesto.html

2 Details of ELSNET available on the website at, http://www.elsnet.org/