1

The Very Long Way from Basic Linguistic Research to Commercially Successful Language Business: the Case of Two-Level Morphology

ANTTI ARPPE

1.1 Introduction

In the Nordic countries, Finland has stood out in the number of start-ups commercializing language technology, as until the late 1990s practically all of the language technology companies founded in the Nordic countries were of Finnish origin. This fledgling Finnish language industry has strong academic roots – a majority of the Finnish IT companies that are primarily involved in creating and providing software products based on language technology can trace their origins to individual researchers or research groups at Finnish universities. But that is where the similarities end. Both in the case of 'older' companies founded in the 1980s and the 'second wave' of the 1990s, the paths and strategies from academic start-ups to commercially functioning corporations have varied substantially. With time, some of these companies have found for themselves clear, profitable niches, but for others the quest still continues. Nevertheless, a major international breakthrough for a Finnish language engineering company is still in waiting. (Arppe 2002)

From research concerning new technology-based startups it is generally known that success is very difficult to predict. A commonly accepted maxim is that out of the twenty start-ups that a venture capitalist invests in, nineteen will at their best barely make even, whereas typically only a single start-up

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

will turn out to be the success story that covers the losses of the others. Even though one can do one's utmost to create an atmosphere which will foster success, one cannot nevertheless control all the external factors in the operative environment, e.g. competitors' actions, national macroeconomic developments, or changes in potential customer expectations, on which the success of any company ultimately hinges. (summarized in Naumanen 2002)

This very same unpredictability and intrinsic riskiness can be said to apply to scientific enterprise. The purpose of this article is to describe how very long, unexpected and winding paths the advancement of science and business can follow by using as a case example the road from basic linguistic research to Kimmo Koskenniemi's dissertation (1983) introducing the two-level model (TWOL), a milestone in computational linguistics, and further on to the final successful commercialization of this model.

1.2 The scientific roots and infant steps of two-level morphology

As Karttunen and Beesley (this volume) outline the individual twists and turns that led to the presentation of two-level morphology by Kimmo Koskenniemi in 1983, these developments will not be discussed in depth in this article. What is worth noting, however, is that the roots of this, in its essence a computational theory are commonly seen to trace back to the sister field of general linguistics, namely to the generative model of the phonological structure of English by Chomsky and Halle in 1968. This seminal work was on its own part a product of a discussion concerning the general modeling of phonetic and phonological structure of any language based on some group of binary distinctive features, initiated by Jakobson, Fant & Halle in 1952.

In conjunction with presenting his theoretical model, Koskenniemi also demonstrated that his approach worked in practice for at least one natural language by implementing the model for Finnish, which was the origin of a software program that was to be later commonly known as FINTWOL. However, there were theoretical doubts as to the general applicability, efficiency and robustness of the two-level model for any given language (Ritchie et al. 1992: 13-39). For instance, Barton (1986) demonstrated that a linguistic description according to the two-level model could in its worst case turn out to be NP-hard. In response to this critique, Koskenniemi and Church (1988) argued that *natural* languages did not exhibit the types of complexity or long-range dependencies which would lead to such computational complexity. For instance, the number of dependencies which were simultaneously in effect over the entire length of orthographical words (and which would thus be a cause of complexity in two-level models) was at its maximum two in natural languages, say vowel harmony in Turkish.

Despite this on-going theoretical debate, researchers started quite rapidly after Koskenniemi's dissertation to apply the two-level model with varying degrees of comprehensiveness for the morphological description of different languages. At the University of Helsinki alone, Finnish was followed by two-level models for Swedish (Blåberg 1984), Swahili (developed in 1985-89 [1], documented in Hurskainen 1992), Akkadian (Kataja & Koskenniemi 1988), French (Norberg 1988), Russian (developed in 1988-1990 [2], documented in Vilkki 1997, 2005), and English (developed in 1989-1990 [3], documented in Heikkilä 1991 and Smeaton, Voutilainen & Sheridan 1990). [4] In this manner, too, the two-level model was demonstrated to work in practice for a wide range of typologically divergent languages with respect to their morphology, whether these languages were predominantly suffixing, prefixing or infixing, or agglutinative or flexional, or long dead or fully alive.

1.3 The commercial potential of the two-level model

The two-level model had obvious practical uses which had great commercial potential in conjunction with software programs for text processing and storage, especially for any European language other than English. The morphology of contemporary English is close to non-existent, and compound words are not written together. Therefore, in the case of English text the major challenge in developing a spell-checker for a word processor is how to compile and compress a comprehensive list of words in the vocabulary, as neologisms are constructed or introduced via borrowing, not only from Latin, Greek and French, but from practically any language of the world that happens to have a suitable word, e.g. *ombudsman* from Swedish and *sauna* from Finnish. Whatever inflection remains can be taken care of with a very limited set of truncation or rewrite rules, e.g. removing the plural marker *-s* from nouns. Likewise, one need not worry extensively on how to cope with inflected forms or how to separate a compound word into its components in the development of search or indexing functionalities for English text data bases.

Contrary to English, most other European languages employ inflection, often likened to performing the function of prepositions in English, though inflection is by no means limited to this grammatical function. It is essential to understand that inflection is more than just a matter of adding one affix after another to base forms, as the root lexeme and the morphemes adjoined to it, in their theoretical, idealized forms, interact with each other, so that the orthographical surface form, i.e. the spelling of an individual root or a morpheme

[1] Personal communication 11.4.2005 from Arvi Hurskainen

[2] Personal communication 11.4.2005 from Liisa Vilkki

[3] Personal communication 11.4.2005 from Atro Voutilainen

[4] N.B. Many of the two-levels models mentioned here have been substantially developed further since these initial versions and their documentation., i.e. the ones for Finnish, Swahili, Russian and English.

always depends on the entire morphological structure of an inflected word. In further contrast to English, common neologisms in many other European languages are to a great extent constructed by using productive mechanisms such as combining existing words in the vocabulary or by derivation, rather than by borrowing. Therefore, in order to perform spell-checking or indexing in inflecting languages, truncation is simply of no practical use. The example below providing several morphological constructions based on the Finnish word *vesi* ‘water’ illustrates this perfectly: [5]

vesi+SG(Singular)+NOM(Nominative):vesi ’water’
vesi+SG+GEN(Genitive):veden ’of water’
vesi+SG+ESS(Essive):vetenä ’as water’
vesi+DN-NEN(Nominal Derivation with *-nen*):vetinen ’watery’
vesi+DN-STO+SG+NOM:vesistö ’water system (group of waters)’
vesi+SG+NOM#pula+SG+NOM:vesipula ’shortage of water’
vesi+SG+GEN#tarve+SG+NOM:vedentarve ’need of water’
vuoro+SG+NOM#vesi+SG+NOM:vuorovesi ’tide (water)’

Thus, to quantify the nature and magnitude of the challenge faced in developing language tools for languages other than English, in e.g. Finnish one can theoretically in the case of the open word classes construct some 2,000 different inflected forms for every noun, 6,000 for every adjective, and 20,000 for every verb. [6]

Should one want to enumerate all the possible inflected forms of, say the 100,000 most common and frequent Finnish words of these open (inflecting) word classes, assuming a distribution as observed in newspaper text, [7] the theoretical sum total would exceed well over 300 million word forms[8]. Even

[5] Notation: MORPHOLOGICAL STRUCTURE:SURFACE FORM; where ‘+’ denotes a morpheme boundary, and ‘#’ denotes a compound boundary

[6] The exact number of morphologically constructible forms is often calculated as 1,872 for Finnish nouns (2 numbers X 13 cases X 6 possessives X 12 clitics) and over 20,000 for Finnish verbs, the latter figure depending on how participle forms are counted in the figure ([530 finite forms + 320 infinitives] X 12 clitics + 5 participles X 1,872). The number of so-called core forms, ignoring clitics, is considerably smaller. Of all of these forms, only a fraction can be observed in even very large corpora of millions of words (personal observations of the author in context of this and earlier work)

[7] This distribution is based on two month’s worth of Helsingin Sanomat, Finland’s major daily newspaper (January and February 1995), available at the Finnish text bank (Helsingin Sanomat 1995) and automatically morphologically analyzed with the Functional Dependency parser for Finnish (FI-FDG) developed by Connexor (Tapanainen & Järvinen 1997). Selecting the base forms in this corpus of approximately 3.2 million unambiguously analyzed running words, all these different inflected forms were found to represent 113,626 common or proper nouns, 12,005 adjectives and 6,641 verbs. On the basis of this, a rough distribution into 86% nouns, 9% adjectives and 5% verbs was established.

[8] 86,000 nouns X 2,000 + 9,000 adjectives X 6,000 + 5,000 verbs X 20,000 = 172M+54M+

with the best compression algorithms there would be no point in trying to generate and list all these forms.

A morphological analyzer program developed according to the two-level model can provide the base form for any inflected word, as long as this word can be constructed using the root lexemes and the morphological rules concerning compounding, derivation and inflection. With these same prerequisites, such an analyzer can also provide the components of any compound word. Furthermore, if a word can be analyzed, provided that the incorporated model is an accurate representation of the orthographical and morphological rules and norms of a language, this will mean that such a word is correctly spelled – in the language in question, that is. Therefore, a two-level model can be used as a basis for the significant improvement of spell-checking and indexing tools for languages with extensive inflection, derivation or compounding. In the case of spell-checking, one needs only to include the root lexeme and its inflectional category in the lexicon in order to recognize not only all the inflected and derived forms of the root but also all the compound words in which it might be used. In the case of indexing and search, one can accurately retrieve all the occurrences of both the inflected forms of a base form and its occurrences as a component of compound words, which is many cases would have been practically impossible or useless with the use of truncation or wildcards. For example, with the two-level model for Finnish, the rather complex but actually observed compound word *väitöskirjatyönohjausajanvarauslista*, i.e.

väittää+DV-OS+SG+NOM#kirja+SG+NOM#...
...työ+SG+GEN#ohjata+DV-US+SG+NOM#...
...aika+SG+GEN#varata+DV-US+SG+NOM#lista+SG+NOM

'reservation list of guidance times for dissertation work' can be correctly recognized in all its inflected forms, e.g. *väitöskirjatyönohjausajanvarauslistallanihan* 'surely on my reservation list for ...', and it can be correctly retrieved using any of its components, e.g. *väitöskirja* 'dissertation' or *aika* 'time' The detection of compound boundaries can also be used to improve hyphenation, as some valid hyphenation borders cannot be detected solely according to character-based rules.

Another commercially interesting property of the two-level model lies in the fact that the model can intrinsically be operated in both directions, i.e. in addition to analysis it can also be used to generate any acceptable morphological form or combination of a root lexemes according to the incorporated linguistic rules. In languages with extensive morphology, this feature turns out to be very useful in the generation of suggestions for corrections of misspelled words, as these correct forms cannot be comprehensively enumerated

100M = 326M word forms

due to the reasons presented above.

The very embodiments of this bidirectional nature of the two-level model are so-called inflecting thesauri, which combine the semantic content of a synonym dictionary or thesaurus with a two-level morphological model for the appropriate language. In these linguistic tools, provided an inflected form as input, the analytical capability of the two-level model is first used to retrieve both the base form and the associated morphological data. Then, the base form is used to retrieve the appropriate synonyms. Finally, the generative capability is coupled with the original morphological analysis data to provide the synonyms of the originally input word in the matching morphological forms.

Nevertheless, one must remember that the two-level model is in the first place a morphological, i.e. structural, rather than a semantic model, and was originally used for linguistic analysis and recognition, in which case the input language is assumed to be orthographically correct. Thus, the recognition of a word does not mean that the recognized word is necessarily a good one in the given context or that it semantically makes any sense – it simply means that the word is morphologically possible. All too often the typos of very common words can be given such a theoretically possible but amusing interpretation, e.g. **ko#mission* ‘cow mission’ instead of *kommission* ‘commission’, or **vis#te* ‘song tee’ instead of *visste* ‘knew’ in Swedish. Likewise, whereas it is very satisfactory both as an end-user and as a developer of a spell-checker to receive *kielitiede* ‘linguistics’ as the only and correct suggestion for the typo **kielittiede*, this is not the case for a slightly different typo, **kielitide*, where one has to sift through six other alternatives, e.g. *?kieli#taide* ‘language/tongue art’, *?kieli#tilde* ‘tongue tilde’, *?kieli#nide* ‘language volume’, *?kieli#kide* ‘language crystal’, *?kieli#side* ‘language/tongue tie’, and *?kieli#tie* ‘language road’, which are either odd or utterly jibberish.

However, rising above these undesirable side-effects observed in the development of practical linguistic software, the generative side of a two-level model can be seen from the perspective of general linguistic theory as a manifestation of the semantic potential of the morphological system of a language that it describes, and its misgivings demonstrate how little of this space a language actually uses. Restricting this undesirably excessive word form generation, derivation and compounding without crippling the system’s openness is the major challenge in developing spell-checkers, and also inflecting thesauri, based on the two-level model. Discussions concerning the practical extent of this inflectional generality as observed in the development of inflecting thesauri for the Scandinavian languages can be found in Arppe et al. 2000 and Arppe 2001.

1.4 The winding path of commercialization

The two-level model for Finnish attracted rapidly the interest of the industry, and Finnish and foreign companies wanted to study, test or use the software for various purposes either as such or desired some form of additional development. This generated a number of commissioned joint projects which employed many researchers at the Department of General Linguistics from time to time. In the 1980s, however, organizing and managing such commercial projects under the auspices of Finnish universities, even at a small scale, was a novel activity, and in contrast to the present there were even less well established forms for it. Furthermore, the general mood in the Finnish academia at the time was that research and business did not mix well, and this was also the conclusion of Koskenniemi and his collaborator, Fred Karlsson, who as head of the Department had to balance the goals and needs of both the basic academic research and teaching activities and the commercial projects at the Department. Therefore, they decided to move these commercial activities to a private company, Lingsoft, which they founded in 1986 (see Knuuttila 2006 for a detailed description and analysis of the views and motivations of the various actors involved at the Department).

For the rest of the 1980s and the early 1990s, this move appears to have had rather an organizational than an economic effect. Researchers who earlier would have worked in the commercial projects at the Department simply continued the same activities at Lingsoft. Koskenniemi took care of the necessary administrative duties as a part-time managing director while continuing as a full-time senior researcher at the Department. The company had no permanent employees nor did it engage in aggressive marketing activities, and people were employed on a case-by-case basis in order to complete some externally commissioned project, or to pursue some research interest of Koskenniemi or Karlsson, which could be financed from the profits of the commercial projects. Sometimes these noncommissioned projects had no direct commercial goal, but would produce resources that would turn out many years afterwards to be of great value by facilitating, accelerating and in some cases simply making possible some later product development efforts. For instance, as two researchers, Katri Olkinuora and Mari Siiroinen, had compiled a synonym dictionary for Finnish in 1989-91, the company did not have to source and license or develop from scratch this resource when it was necessary in order to develop a Finnish inflecting dictionary for Microsoft in 1995-6.

During this initial, project-driven phase of Lingsoft, the annual turnover of the company hovered on the average at just below one hundred thousand euros. However, at the same time the company succeeded in closing some individual deals, which by themselves even exceeded the average annual turnover. The most important of such deals was the licensing of the Finnish spell-checker and hyphenator to WordPerdfect in 1988, and the Finnish base form indexing as part of the article data base of Helsingin Sanomat, Finland's largest daily newspaper, in 1992.

In 1992, the persevering efforts at the Department to turn computational linguistics into its own independent discipline bore fruit, as the chair of computational linguistics (renamed language technology in 1999) was established permanently at the University of Helsinki, with Koskenniemi as its first holder. It was then that a slow transformation into a commercial company operating in the traditionally understood sense began at Lingsoft, which was marked by the hiring of the first permanent employee, Krister Lindén, as the managing director. The company embarked on its first major technological development project, undertaken entirely by the company, in order to develop a two-level model for German. This project culminated in 1994 in the overall victory of the first German Morpholympics (Hausser 1996), a competition on developing an efficient and comprehensive morphological analyzer for German in which Lingsoft with its GERTWOL (Koskenniemi & Haapalainen 1994) was the only commercial and non-German participant.

However, this victory did not immediately produce economic returns which had been invested in it, as one might have expected based on the size of the German market and the enthusiastic commercial reception experienced earlier with regards to FINTWOL. Other external developments were also presenting rising challenges for the company. Lingsoft's long-standing partner and customer of proofing tools [9], WordPerfect, was increasingly losing market share in its main business of word-processors to Microsoft, which could leverage its dominance in the operating systems market. Microsoft, on the other hand, already had an existing licensing deal for all its proofing tools for all the major European languages with Inso [10], which had transformed from the spin-off software division of the American publisher Houghton Mifflin into the major player in the language industry in the earlier 1990s. Though Inso's proofing tools were in essence word-list-based, following the English model as presented above, Microsoft was apparently under no sufficient customer pressure to change its subcontractor in 1994-5. [11]

[9] The term Proofing Tools has become to denote not only text-verification programs such as spell-checkers and grammar-checkers also hyphenators and thesauri, i.e. synonym dictionaries, mainly as a result of the influence of the major licensors of these tools, firstly WordPerfect and later Microsoft.

[10] The company in question has operated under several company names, first as InfoSoft International Incorporated 1994-1995, then as Inso Corporation 1995-2000, and finally as eBT International 2000-, being liquidated in 2001. In 1998 the company, then as Inso, sold off its entire linguistic tool business, including customer relationships and contracts, to Lernout&Hauspie, itself now also defunct.

[11] Personal communications in October 1994 and 14.11.1995 from Tarja Tiirikainen, Program

Therefore, from Lingsoft's perspective Microsoft seemed to be a lost cause at that time, and the only available path to generate revenues from proofing tools would be to aim directly at each national end-user market. On the other hand, it appeared that the window for proofing tools as independently marketed software was closing, based on the competitor analysis of e.g. Kielikone, another Finnish language technology start-up, which seemed to have been shifting its marketing and development focus away Morfo, its reputed stand-alone spell-checker for Finnish, to electronic dictionaries. Despite some initial distaste to ‘annoying squiggly red lines” marking typos, later developments have shown that proofing tools indeed have turned into the deeply embedded and enabling components of other software programs (EUROMAP 1998: 17-20), the functionality and quality of which are only indirectly visible to the end-users of the parent applications such as word-processors. In conjunction, the structure of the supply chain for proofing tools has developed into a true niche-channel supplier model, with Lingsoft and other small language technology companies in the role of niche suppliers, and Microsoft and other international IT giants as the channels (EUROMAP 1998: 47-56).

Though Microsoft was thus not overtly interested in relicensing its proofing tools, it was interested in localizing AnswerWizard, a fusion of a natural language database query system with a help database, and it was keen on having this work undertaken by a company with linguistic technological competence. Lingsoft was obviously such a company, with demonstrated experience in a variety of languages. However, the range of languages offered to Lingsoft were not only those in which the company had previous experience of its own or through partnerships, such as Swedish and Danish, but also languages with which the company had no real previous competence, such as Norwegian, Dutch and Spanish. Nevertheless, Lingsoft succeeded in negotiating in 1995 a deal covering the localization of AnswerWizard for all of the mentioned languages. Even more importantly, Lingsoft also satisfied Microsoft's quality and other requirements, as the project was renewed, with the addition of Finnish, Russian, Czech and Polish, on several occasions until 2000.

Not only was the AnswerWizard project instrumental in providing Lingsoft desperately needed financial stability in 1995-1996, but by demonstrating the company's capability to undertake such a demanding and complex multlilngual project it also put Lingsoft in a favorable position when Microsoft finally, and in fact quite soon, did decide to reconsider its proofing tool licensing relationships. Thus, Lingsoft had the full package of sufficient financing, right contacts, and good track record, in addition to its birthright of state-of-the-art linguistic technology, in order to be selected in 1996 out

manager for proofing tools at Microsoft.

of three competitors as Microsoft's new subcontractor for the Finnish spell-checker, hyphenator and thesaurus, which relationship Lingsoft has retained with Microsoft ever since. It is a tribute to Koskenniemi's linguistic skills to note that the linguistic description of Finnish incorporated by him in FINTWOL was to a very large extent used in its original form in these proofing tools licensed to Microsoft, and this still continues to be very much the case, even after over twenty years of their original inception.

After this suite of Finnish proofing tools, Lingsoft went on to license to Microsoft the Swedish inflecting thesaurus in 1996, the Swedish spell-checker and hyphenator and the Norwegian (bokmål) and Danish inflecting thesauri in 1997, and the German spell-checker, hyphenator and inflecting thesaurus, and the Norwegian (both bokmål and nynorsk) spell-checkers and hyphenators and the Danish spell-checker and hyphenator in 1998. In addition to these proofing tools based essentially on the two-level model, Lingsoft also succeeded in developing in 1997-1998 and licensing to Microsoft a Swedish grammar-checker (Arppe 2000, Birn 2000), the first of its kind, which was based on the Constraint Grammar formalism originally presented by Fred Karlsson (1990), and realized and further developed by the Research Unit for Multilingual Language Technology (RUMLAT) (Karlsson, Voutilainen, Heikkilä & Anttila 1995). This product development process was successfully duplicated for Finnish, Danish and Norwegian (bokmål), and licensed to Microsoft in 2000-2001. In association with these successful contracts, the number of personnel and the turnover of the company started to grow as presented in Table 1.

1.5 Factors which influenced the commercialization process

From the introduction of the two-level model in Koskenniemi's dissertation in 1983 it took over ten years to transform this theory into a steady commercial income flow of over one million euros in 1996, if measured in terms of Lingsoft's annual turnover presented in Table 1. With the benefit of hindsight one can consider whether it would have been possible to significantly accelerate this process of technology transfer and commercialization.

The basic building blocks used by Lingsoft in its proofing tools, e.g. the two-level models for Finnish and Swedish, had been extensively developed by 1990, which is demonstrated by the licensing deal of a Finnish spell-checker to WordPerfect as early as in 1988. The inhibiting factors were essentially technical in nature and intrinsic to the initial development and implementation of the two-level model. At the University of Helsinki, Koskenniemi had at his disposal the best and most advanced computer facilities available to anyone in Finland, which already by the beginning of the 1980s used multiprogramming operating systems with virtual memory. As a component of a

FIGURE 1 Lingsoft's turnover and personnel 1992-2005.

Year	Turnover	Personnel
1992	0.4 MFIM (0.06 M€)	2
1993	1.3 MFIM (0.22 M€)	5
1994	1.4 MFIM (0.23 M€)	6
1995	4 MFIM (0.7 M€)	7
1996	6 MFIM (1.0 M€)	10
1997	8 MFIM (1.3 M€)	16
1998	13 MFIM (2.2 M€)	20
1999	8 MFIM (1.3 M€)	25
2000 (15 months)	17 MFIM (2.8 M€)	30
2001	1.2 M€	60
2002	0.15 M€	10
2003 (Pasanet merger)	0.25 M€	4
2004 (estimate)	0.08 M€	10
2005 (estimate)	1.9 M€	15

functioning computer program, the obvious data structure into which the two-level model could be transformed was a single finite-state automaton, which in the case of the original FINTWOL consumed several hundred kilobytes of memory, competing with the memory needs of other applications. This had not been a problem in the computing facilities at the university. However, in the realm of personal computers which were the source of commercial potential for general software programs, as Microsoft's MS-DOS and Windows operating systems had gained a dominant position by the beginning of 1990s, operating systems that in practice allowed for running multiple applications concurrently, with genuinely flexible and sufficient virtual memory, were to spread broadly on the general consumer market only with the introduction of the Windows95 in 1995. In principle, it would have been possible to hack a solution to make two-level models (of the full-fledged size necessary for e.g. spell-checking) work with the memory constraints of earlier applications and operating systems – certainly the existence of Kielikone's Morfo spell-checker was a practical proof of its feasibility – but with the limited personnel resources at the disposal of the company it was simply not considered worth all the effort, as the inevitable arrival of Windows95 was more or less certain for several years before its eventual launch. [12] At the time, these arguments seemed from the perspective of Lingsoft's marketing personnel some sort of

[12] Personal communications with Pasi Ryhänen in 18.3.2005 and Mikko Silvonen on 18.3.2005, who were both senior software engineers and product development managers at Lingsoft throughout the 1990s.

unfounded resistance or reluctance of the software developers, but now it is quite clear that in 1994 or even in 1993 the necessary development investments would not have been offset by any potential benefits and associated incomes in the year or two by which Windows95 had effectively replaced earlier PC operating systems.

Even without this intrinsic technical restriction on the spread of linguistic tools exploiting two-level models, the embedded nature of proofing tools had the logical consequence that no proofing nor other purely linguistic tools would have had any commercially interesting market potential before the spread of word-processors or text data bases, which has been described as the product adoption hierarchy of linguistic tools (Arppe 1995a, 1995b). Without software programs that enabled the electronic authoring of text there would hardly have been any commercial need for software programs to spell-check such texts. Even more fundamentally, electronic text authoring tools could become general household consumer tools only with spread of personal computers, which started in earnest with the introduction of IBM's first PCs in the early 1980s – at the very same time that the two-level model was conceived of in the first place. Therefore, it is difficult to see how proofing tools which capitalized on the rule-based, open nature of the two-level model, allowing for a crucial improvement when compared with the preceding list-based solutions, could in practice have been successfully commercialized essentially earlier than how the events folded out in practice.

1.6 Conclusions on the nature of scientific and commercial advancement

In conclusion, Koskenniemi's two-level model in 1983 was a practical computational solution to originally linguistic research questions and subsequent discussion which can be traced as far back as 1968, and even earlier. On its own part, the commercialization process of the two-level model was greatly dependent on developments in the external IT business environment. The introduction of the first personal computers in the early 1980s and then the spread of word-processors and text data bases in the late 1980s were obligatory prerequisites for the emergence of a need for linguistic proofing tools. The break-through of such solutions based on the two-level model was further dependent on the large-scale spread of 32-bit operating systems starting in the mid 1990s.

With the help of these external developments and as a result of all the development work at Lingsoft in 1986-2001, the company became Microsoft's subcontractor of proofing tools for all the major Nordic languages and German. Despite severe difficulties experienced by the company in 2001-2004 (see Knuuttila 2006 and Arppe 2002), finally relieved by a merger with

Pasanet, a translation and localization company based in Turku, these contracts have all been renewed again in 2004, and in retrospect proofing tools can clearly be seen to have in practice been for Lingsoft both its core technological competence and its main source of income over its entire existence. Thus, Finnish language technology, based on the two-level model and the constraint grammar formalism, is now used by tens of millions of people in the Nordic and German-speaking countries, which can be considered a major success for the Finnish IT industry, and even more so for the Finnish language technology community. In this, Kimmo Koskenniemi has played a central role.

In order for all this to be possible we can see an overall arch of incremental individual advances in basic research spanning over several decades from the 1950s to the 1990s. It is clear that the present, but by no means final scientific and commercial outcomes could not have been predicted or determined at the outset. The history leading to and proceeding on from the two-level model is an outstanding example of how scientific research can produce significant commercial benefits, when it is allow to proceed in a free and open manner, and is not constrained by any short-term interests whatsoever.

1.7 Acknowledgements

The historical accounts in this article are based on interviews with Kimmo Koskenniemi and others conducted for my Master's thesis, which Kimmo arranged for me to undertake at Lingsoft in 1994-1995. I furthermore gratefully acknowledge the innumerable discussions with past and present employees at Lingsoft who participated in the development of the company's proofing and other tools, whether the subject was corporate strategy or language technology. Specifically, I would like to thank Thomas Bilgram, Juhani Birn, Era Eriksson, Juhani Järvikivi, Risto Kankkunen, Krister Lindén, Anni Koskenniemi, Ilkka Koskenniemi, Timo Koskenniemi, Ari Majorin, Mariikka Majorin, Sjur Nørstebø Moshagen, Markku Norberg, Ari Paavilainen, Otto-Ville Ronkainen, Pasi Ryhänen, Mikko Silvonen, Eero Taipale, Kimmo J. Tykkälä, Eleonoora Vanhanen, Mari Voipio, Fredrik Westerlund, and Malene Würtz, for being part in realizing the potential of what the two-level model and Lingsoft could be. With regards to this article, I am grateful to Martti Vainio and Anssi Yli-Jyrä for providing input concerning the phonological, computational and philosophical roots of the two-level model, and to Krister Lindén for comments on its commercialization. I am also indebted to Jussi Piitulainen and Kimmo Koskenniemi for encouraging me in 1999 to set up of a lecture course, on which the analysis presented here is originally based upon. Finally, I would like to thank Tarja Knuuttila for providing perspective which I believe has led a more balanced account of the events.

References

Arppe, A. 1995a. *The Strategic Opportunities of a Small High-Technology Company on Emerging Markets.* Unpublished Master's Thesis, Institute of Industrial Management, Helsinki University of Technology.

Arppe, A. 1995b. Information Explosion and the Use of Linguistic Tools in Finland. Harakka, T. & Koskela, M. (eds.) *Kieli ja tietokone.* Association Finlandaise de Linguistique Appliquée, Yearbook 54. Jyväskylä.

Arppe, A, Voipio, M. & Würtz, M. 2000. Creating Inflecting Electronic Dictionaries. Lindberg, Carl-Erik & Lund, Steffen Nordahl (eds.) 17th Scandinavian Conference of Linguistics, Nyborg August 20-22, 1998. *Odense Working Papers in Language and Communication* No 19, April 2000, Vol 1. University of Southern Denmark, Odense, Denmark.

Arppe, A. 2000. Developing a Grammar Checker for Swedish. In: Nordgård, T. (ed.) *Proceedings from the 12th Nordiske datalingvistikkdager*, Trondheim, December 9-10, 1999. Department of Linguistics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

Arppe, A. 2001. Lärdomar från utveckling av inflektereande synonymordböcker. Gellerstam, Martin et al (eds.) Nordiska studier i Lexikografi 5. Rapport från *'Konferens om lexikografi i Norden'*, Gothenburg May, 26-29, 1999. Skrifter utgivna av Nordiska föreningen för lexikografi (6) i samarbete med Nordiska språkrådet och Meijerbergs institut, Gothenburg, Sweden.

Arppe, A. 2002. Ei yhtä ainoaa polkua - Suomalaisia kokemuksia matkalla kieliteknologisesta tutkimuksesta liiketoimintaan [No single path - Finnish lessons in the commercialisation of language engineering research]. *Puhe ja kieli* 22:1, pp. 37-44. English translation available at: URL: http://www.hltcentral.org/page-969.shtml.

Barton, E. 1986. Computational complexity in two-level morphology. In: *Proceedings of the 24th Conference of the Association for Computational Linguistics*, pp. 53-59.

Birn, J. 2000. Detecting grammar errors with Lingsoft's Swedish grammar-checker. In: Nordgård, T. (ed.) *Proceedings from the 12th Nordiske datalingvistikkdager*, Trondheim, December 9-10, 1999. Department of Linguistics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

Blåberg, O. 1984: *Svensk böjningsmorfologi. En tvånivåbeskrivning*. [The inflectional morphology of Swedish. A two-level model description]. Unpublished Master's thesis, Department of General Linguistics, University of Helsinki.

Chomsky, N. & Halle, M. 1968. *The Sound Patterns of English.* New York: Harper and Row.

EUROMAP 1998. *The EUROMAP Report. Challenge & Opportunity for Europe's Information Society*. Language Engineering Sector of the Telematics Applications Programme, DG XIII Telecommunications, information market and exploitation of research, European Commission, Luxembourg.

Hausser, R. 1996. *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*. Max Niemayer, Tübingen.

Heikkilä, J. 1991. *A Lexicon and Feature System for Automatic Morphological Analysis of English.* Unpublished Master's thesis, Department of General Linguistics and Department of English, University of Helsinki.

Helsingin Sanomat 1995. 22 million words of Finnish newspaper text. Compiled by the Department of General Linguistics, University of Helsinki and CSC Tieteellinen Laskenta Oy. Available at URL: http://www.csc.fi/kielipankki/

Hurskainen, A. 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology. An Application to Swahili. *Nordic Journal of African Studies*, Vol 1/1. Available at URL: http://www.njas.helsinki.fi/.

Jakobson, R., Fant, G. & Halle, M. 1952. *Preliminaries to speech analysis: the distinctive features and their correlates.* (MIT Acoustics Laboratory Technical Report 13.) MIT Press, Cambridge, Massachusetts.

Karlsson, F. 1990. Constraint Grammar as a Framework for Parsing Unrestricted Text. In: Karlgren, H. (ed.), *Proceedings of the 13th International Conference of Computational Linguistics*, Vol. 3. Helsinki 1990, 168-173.

Karlsson , F., Voutilainen, A., Heikkilä , J. & Anttila, A. 1995. *Constraint Grammar: A Language-Independent Framework for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin/New York.

Karttunen, L. & Beesley, K. R. 2005 (this volume). Twenty-five years of finite-state morphology.

Koskenniemi, K. 1983. *Two-level morphology: A general computational model for word-form recognition and production.* Publication 11, University of Helsinki, Department of General Linguistics, Helsinki.

Koskenniemi, K. & Church, K. W. 1988- Two-level Morphology and Finnish. In: *Proceedings of the 12th International Conference on Computational Linguistics*, COLING-88, Budapest, ed. D. VARGHA, 1, pp. 335-339, John von Neumann Society for Computing Sciences, Budapest (1988). B2U.

Koskenniemi, K. & Haapalainen, M.. 1994. GERTWOL: Ein System zur automatischen Wortformerkennung deutscher Wörter. In: Hausser, R. 1996. *Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994*, pp. 121-140. Max Niemayer, Tübingen. Also available at URL http://www.lingsoft.fi/doc/gertwol/intro/gertwol.txt.

Kataja, L. & Koskenniemi, K. 1988, Finite-state Description of Semitic Morphology: A Case Study of Ancient Accadian. In: *Proceedings of the 12th International Conference on Computational Linguistics, COLING-88*, Budapest, ed. D. VARGHA, 1, pp. 313-315, John von Neumann Society for Computing Sciences, Budapest (1988). B2U.

Knuuttila, T. (forthcoming in 2006). Harmaalla alueella Monikielisen kieliteknologian yksikkö tutkimuksen ja kaupallistumisen ristipaineessa. In: *Yliopistotutkimuksen muutos ja tietoyhteiskunnan sisäinen ristiriita* [The transformation of academic research and the internal paradox of the information society], Helsinki University Press, Helsinki.

Naumanen, M. 2002. *Nuorten teknologiayritysten menestystekijät* [Success factors of new technology-based companies] (Sitra reports series, ISSN 1457-571X; 28). Edita Publishing Oy, Helsinki. ISBN 951-37-3819-1

Norberg, M. 1988: *Koskenniemen kaksitasomallin mukainen, foneemipohjainen kuvaus ranskan adjektiivien ja substantiivien taivutusmorfologiasta* [A phoneme-based description of the inflectional morphology of French adjectives and nouns according to Koskenniemi's two-level model]. Unpublished Master's thesis, Department of General Linguistics, University of Helsinki.

Ritchie, G., Russell, G., Black, A. & Pulman, S. 1992. *Computational Morphology: Practical Mechanisms for the English Lexicon*. MIT Press, Cambridge, Massachusetts.

Smeaton, A. F., Voutilainen, A. & Sheridan, P. 1990. The application of morphosyntactic language processing to effective text retrieval. In: *Esprit '90 Conference Proceedings*, pp 619-635, Dordrecht.

Tapanainen P. and Järvinen T. 1997. A non-projective dependency parser. Proceedings of the 5th *Conference on Applied Natural Language Processing* (ANLP'97), pp. 64-71. Association for Computational Linguistics, Washington, D.C.

Vilkki L. 1997. *RUSTWOL: A System for Automatic Recognition of Russian words.* Available at URL: http://www.lingsoft.fi/doc/rustwol/

Vilkki L. 2005 (this volume). RUSTWOL: A Tool for Automatic Russian Word Form Recognition.