OSLa
Oslo Studies in Language
3 (2) / 2011

Janne Bondi Johannessen (ed.)

Language Variation Infrastructure

Papers on selected projects

Oslo Studies in Language
General editors: Atle Grønn and Dag Haug

OSLa
Oslo Studies in Language
3 (2) / 2011

Janne Bondi Johannessen (ed.)

Language Variation Infrastructure

Papers on selected projects

Oslo Studies in Language, 3(2), 2011.
Janne Bondi Johannessen (ed.):
Language Variation Infrastructure. Papers on selected projects.
Oslo, University of Oslo

ISSN 1890-9639

Set in LaTeX fonts Gentium Book Basic and Linux Libertine by
Rune Lain Knudsen, Vladyslav Dorokhin and Atle Grønn.
Cover design by UniPub publishing house.
Printed by UniPub from camera-ready copy supplied by the editors.

`http://www.journals.uio.no/osla`

Contents

J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 1–4. (ISSN 1890-9639)
http://www.journals.uio.no/osla

INTRODUCTION

JANNE BONDI JOHANNESSEN
Department of Linguistics and Nordic Studies, University of Oslo

Research infrastructure has become more and more important in recent years, not just in the natural sciences, but also in the humanities and especially in linguistics. While linguists over the last fifty years have often found it sufficient to use their own intuitions about grammaticality by way of introspection when making claims about language structure, such an approach is rarely seen as acceptable in the 21st century. With electronic text corpora and even speech corpora the linguists have far better opportunities for making sound claims about empirical matters, and, importantly, in a way that can be scientifically checked by other researchers. In the study of language variation, it is even more important to have access to resources that document this variation.

This volume was initiated after a workshop on Research Infrastructure for Language Variation Studies (RILIVS), held at the University of Oslo in the autumn of 2009. The workshop was the main one of a series of three exploratory workshops funded by the Nordic research fund NOS-HS and initiated by researcher Øystein Alexander Vangsnes (University of Tromsø) with professor Janne Bondi Johannessen (University of Oslo), in addition to the following people: professor Odd Einar Haugen (University of Bergen), professor Anders Eriksson (University of Gothenburg), professor Höskuldur Thráinsson (University of Iceland), professor Jan-Ola Östman (University of Helsinki), assistant professor Karen Margrethe Pedersen (University of Copenhagen), assistant professor Henrik Jørgensen (University of Aarhus), professor Christer Platzack (University of Lund), researcher Bruce Morén (University of Tromsø), professor Tor A. Åfarli (Norwegian University of Science and Technology). The major aims of the workshop can be summarised thus:

- Bring together scholars and technicians from various projects concerning linguistic variation in order to:
  - consolidate the resources developed in the individual projects
  - establish common and realistic research goals in the field of linguistic variation
  - develop common standards and tools to enhance the quality of linguistic variation research

- Goals to be identified should be targetted at the needs and desires of the linguists and not delimited by the current state of technology. Taking this general stand will foster innovativeness rather than mere application of existing resources.

This volume consists of papers on topics of essentially two types: on the use of language research infrastructure and on development of such infrastructure.

Patrick Bye is interested in linguistic innovations and describes how mapping these geographically can help reconstructing the history of certain linguistic features. He argues that for a particular linguistic feature it may be difficult to determine the order of development just by using simplified generalisations. Instead, important insight can be gained by looking at the actual geographical distribution of a feature, taking into account factors such as difficult terrain, knowledge about the history of other linguistic features, and also the cultural, archeological and historical context of the relevant geographical area.

Harald Hammarström and *Sebastian Nordhoff* focus on the problem of proliferation of linguistic documents and references. In particular, they describe LangDoc, a project for the creation of a website for bibliographical information on linguistic typology. There can be no doubt as to the usefulness of this kind of annotated database, since existing ones are either skewed w.r.t. the authors or types of literature that they represent, or cover only a subset of the languages of the world, or are expensive.

Janne Bondi Johannessen's paper presents the Nordic Dialect Corpus, which is a joint infrastructure where linguists from all the Nordic countries have contributed with dialect recordings, which have subsequently been transcribed, and annotated and made available in a user-friendly searchable corpus. The corpus thus consists of spoken texts from five languages, and is searchable w.r.t. words and parts of words, annotations, phonetic form, and informant metadata. The data are presented in a multimedia display for transcriptions, audio and video, and there are many further options for frequency, sorting, translations and even distribution of search results on geographical maps.

Jan Pieter Kunst and *Franca Wesseling* describe the European Edisyn project, which has a two-fold linguistic goal: on the one hand to establish a network of dialect linguists who have the same standards w.r.t. methodology and data collection, and on the other to study a particular phenomenon: doubling. However, in order to be able to better study the linguistic variation in the languages affiliated with the project, an important outcome is the development of a common database: The Edisyn Search Engine. The database so far consists of five corpora, covering Dutch, Estonian, Northern Italian, Portuguese, and Scandinavian dialects (incidentally the Nordic Dialect Corpus, which is also represented in this anthology). The prospect of a common search engine across languages as different as those mentioned is an exciting one for future linguistic research.

Therese Leinonen focusses on tools for illustrating dialect variation, and in particular, dialect levelling in Swedish. She has used phonetic dialect data available from a previous dialect project, and put them through first acoustic analysis and then dialectometrical methods, in particular aggregating techniques like cluster analysis and multidimensional scaling. The various maps that are projected on the basis of the calculations are very good for illustrating the status of the variation. By using variables such as the age of informants in addition to geographical site it is possible to see how a dialect is changing even if the data have all been recorded at the same time. The changes are very obvious once they are projected in the whole colour spectrum.

Eiríkur Rögnvaldsson, *Anton Karl Ingasson* and *Einar Freyr Sigurðsson* present an ongoing project which aims at finding methods for devloping a Icelandic treebank. Their paper stresses the importance of using open source software, especially for languages like Icelandic, which can be considered to be less-resourced. Their project benefits from the open source tool-kit IceNLP, which has two kinds of morphological taggers, a shallow parser, a lemmatiser, and a tokeniser. Combined with the two available programs Corpus Search and CorpusDraw, the way to a treebank is relatively short.

Diana Santos presents the Linguateca infrastructure for Portuguese. Linguateca is a Portuguese project that has existed for many years. The resources have not been presented before in full for an international audience, and this is the main purpose of this paper. First of all there is a cluster of corpora containing various text genres, written in Portuguese from three continents. The texts have been grammatically tagged and parsed. Parts of the resources have also been semantically classified using automatic methods.

Xavier Villalba's paper is on the Hispacat database of syntactic constructions and how to apply it to syntactic research. This database contains data suitable for micro-comparative variation studies, and contains both Catalan and Spanish data. Importantly, in contrast to an ordinary text corpus, it also contains negative data. The database has three main goals: to be an empirical database for bilingualism and L2 learning, to provide a comparative grammar, and to contain a catalogue of examples.

Without proper peer reviewing academic work would suffer. We are therefore very grateful to the reviewers who accepted to do this job for us:

Sturla Berg-Olsen, Antonio Fabregas, Jan Terje Faarlund, Kristin Hagen, Janne Bondi Johannessen, Karl-Gunnar Johansson, Terje Lohndal, and Andrew Nevins.

WORKSHOP ON RESEARCH INFRASTRUCTURE FOR LINGUISTIC VARIATION:
http://www.tekstlab.uio.no/rilivs/

Speakers and participants at the RILIVS workshop at the University of Oslo 2009. (Photo: Ram E. Gupta)

J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 5–29. (ISSN 1890-9639)
http://www.journals.uio.no/osla

MAPPING INNOVATIONS IN NORTH GERMANIC WITH GIS

PATRIK BYE
University of Tromsø

ABSTRACT

The mapping of innovations, as opposed to taxonomic features, has so far been little used in historical linguistics and dialect geography. Here I show with two examples from Peninsular North Germanic how linguistic theory may cast light on complex mosaics of geographically competing features and how dialect geography can help choose between competing reconstructions. This research builds on a database of more than 50 innovations and over 1000 municipalities in the Nordic countries coupled with mapping software (ArcGIS).

[1] INTRODUCTION

Dialectology has traditionally relied on maps showing the areal distributions of features. [1]

Such maps provide a great deal of useful information for undertaking reconstructions in historical phonology. Although dialectometry makes use of state-of-the-art mapping techniques, historical linguistics has not yet attempted to use areal distribution data to support or refute reconstructions. What I hope to show here is that mapping the spatio-temporal structure of linguistic variation can provide important and, at times, crucial validation of our reconstructions.

To this end, a couple of years ago, I started compiling a database of cultural and linguistic innovations in North Germanic. The database, stored simply as an Excel spreadsheet, currently has data for over 1000 municipalities in Scandinavia and the Nordic region, and covers over 50 innovations. Each municipality is marked as either having the innovation (1) or not (0). The data can be presented visually using mapping software such as ArcGIS (http://www.esri.com/software/arcgis/index.html) in a form that can be adaptable to needs of various users.

[1] The thesis presented in this article has taken shape over several years, during which time I have had the privilege of fruitful discussions with a number of scholars on the research reported here. In particular, I would like to thank Gjert Kristoffersen, Aditi Lahiri, Ove Lorentz, Bruce Morén-Duolljá, Tomas Riad, Curt Rice, and Øystein Vangsnes.

The data has been compiled from traditional maps, such as those in Brøndum-Nielsen (1927), Christiansen (1969), Haugen (1976), Sandøy (1996). These maps are 'taxonomic' in the sense that they represent the areal distributions of *features*; they do not assume a theory of the *innovations* that gave rise to them. It may be difficult to discern the pattern of innovations that has given rise to the feature mosaic represented on the map. In order to explain the feature mosaic we can turn to (i) linguistic theory as a rich source of hypotheses about which innovations are likely and how they might interact sequentially, and (ii) to theories dealing with the diffusion of innovations (Rogers 2003), in particular spatial diffusion (Hägerstrand 1967).

There is, to be sure, an extensive tradition in dialectology dealing with the interpretation of the geographical distribution of dialect features. From the configuration of certain isoglosses[2], for example, it is possible to see that linguistic features spread out from particular centres. An important early contribution in this vein was Kranzmayer (1956), who was able to infer that the Central Bavarian area between Munich and Vienna including the Danube valley was the centre of several phonological innovations in southern German. The spreading of innovations may leave more conservative zones, called relic areas, unaffected. Innovations may also differ in their areal extent. These differences may be explained by the chronology and changes in communication patterns in the network over time.

One important line of inquiry is the work of the Neolinguistic school (Bàrtoli 1925, 1945; Bàrtoli & Bertoni 1925; Bonfante 1947), whose contributions are also discussed by Petyt (1980) and Trudgill (1975). Matteo Bàrtoli and his colleagues attempted to define areal norms which could be used to establish the relative age of geographically competing variants. The most important such norms are listed below in (1). These formulations are adapted from Trudgill (1975, 236).

(1) Given linguistic forms A and B,
  a. If A is found in isolated areas, and B in areas more accessible for communication, then A is older than B.
  b. If A is found in peripheral areas and B in central areas, then A is older than B.
  c. If A is used over a larger area than B, then A is older than B.

In what follows, we'll apply these norms to two examples from North Germanic.

[2] In an innovation-oriented perspective, an isogloss is a line marking the maximum extent of the spread of some innovation, like the line left on the sand after the wave recedes.

[2] STRONG FEMININES IN MID AND WEST NORWEGIAN

A good example of how recoding traditional feature mosaics in terms of innovations and visually interpreting the result comes from the areal distribution of allomorphs of the definite suffix in the strong declension of feminine nouns in varieties spoken in Western and Mid-Norway. Nouns in the strong declension typically have a stem that ends in a consonant, such as *kløv* 'packsaddle', or *bygd* 'country settlement, township'; those in the weak declension end in an unstressed /e/ or /a/, e.g. *ferje* 'ferry'. For further details on this classification, see for example Beito (1986 [1970], 112ff.).

Figure 1 is an adaption of a traditional map from the collection in Christiansen (1969). As can be seen, it presents an initially bewildering mosaic of geographically competing variants, with realizations of the strong feminine definite singular suffix ranging from {-i}, through {-ei}, {-e}, {-æ}, {-ɑ}, and {-ɔ} to {-u}. Trying to apply Bàrtoli's norms directly to Figure 1 to determine the relative ages of the variants does not get us very far. Fortunately, in this case, we can be reasonably sure how this variation arose, thanks both to the evidence of manuscripts and phonetic theory. In Old Norse, the feminine definite suffix had the form *-in* (see e.g. Gordon 1981). In most of Scandinavia, the final *-n* was lost, but a trace of it remained as nasalization on the vowel. This has disappeared in most varieties, although nasalized suffixes still survive in some relic areas. It can still be heard in Älvdalen (Levander 1909) in Dalecarlia, and was still found in Selbu, in South Trøndelag, a century ago (Gjert Kristoffersen *voce*). Nasalization is known to affect vowel quality in various ways. Delvaux et al. (2002), for example, found that nasalized vowels in French had generally lower F2 (had a 'darker' timbre) than their oral counterparts. This qualitative difference may be enhanced by making other changes to the articulation of the vowel, including retracting the tongue body, rounding the lips, lowering (in front vowels), and raising (in back vowels). All of these strategies serve to mimic or reinforce the qualitative effects of nasalization in that they lower F2. We can note that the first eight cardinal vowels, [i e ɛ a ɑ ɔ o u], de-

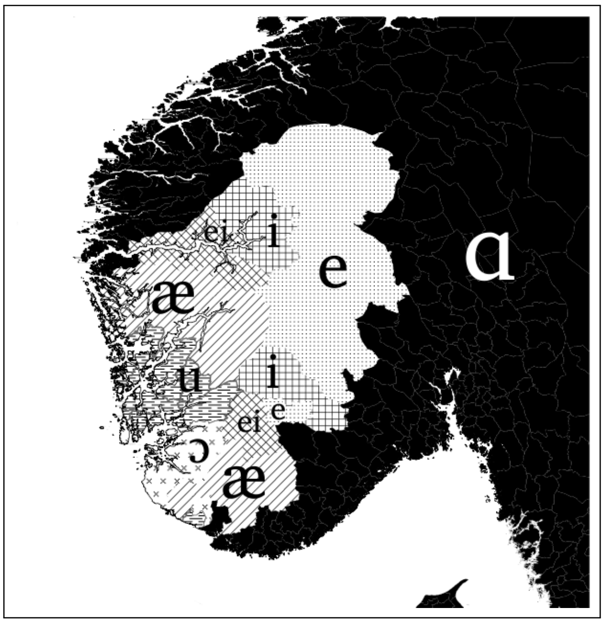


FIGURE 1: Areal distribution of strong feminine definite suffixes in Western Norway

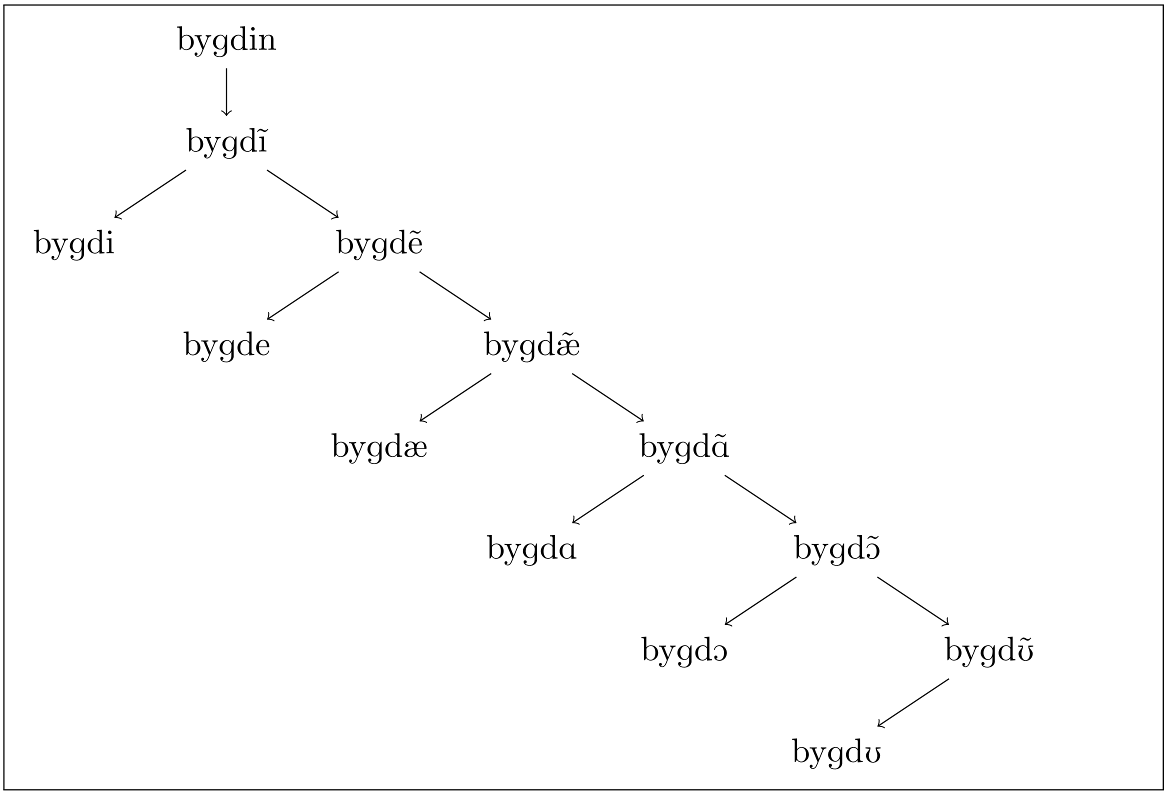


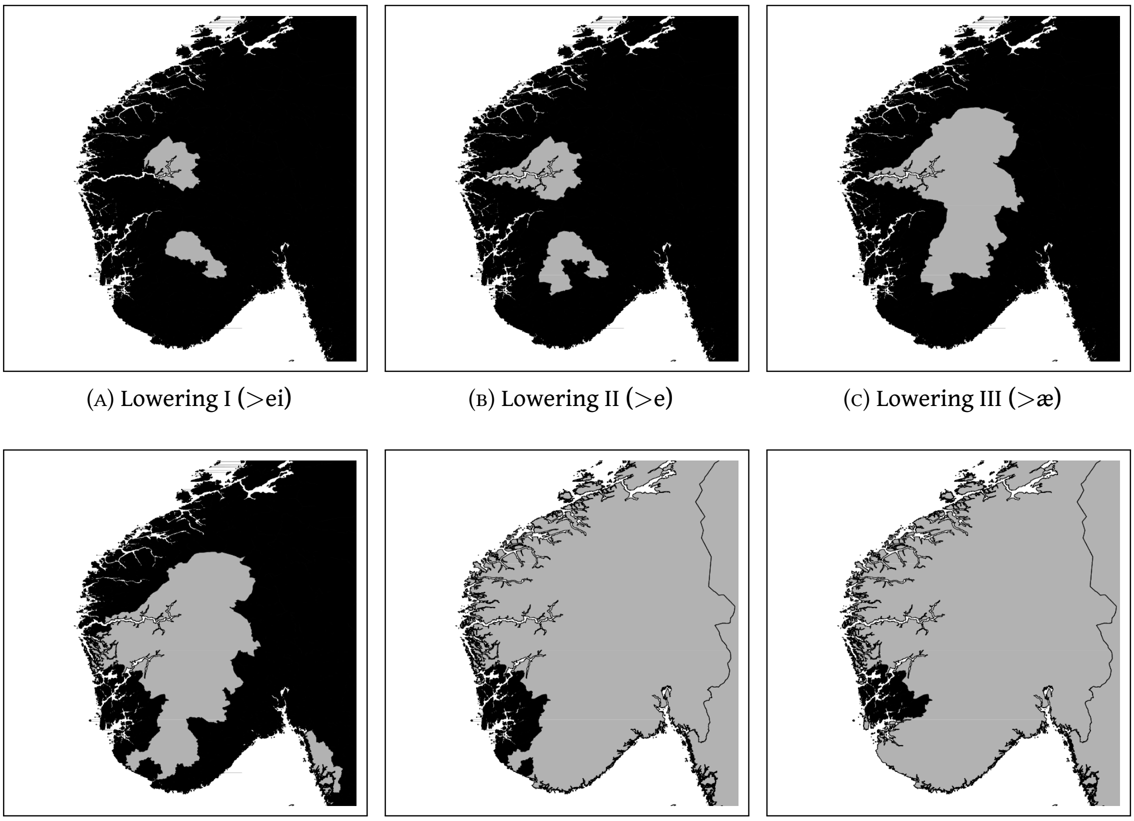
FIGURE 2: Evolutionary pathways in the development of strong feminine definite suffixes in Western Norway

scend in F2.[3] Given this, it is possible to reconstruct a series of vowel quality alterations, beginning with lowering from nasalized [ĩ] through [ẽ] to [æ̃], followed by retraction to [ɑ̃], rounding to [ɔ̃] and, finally raising to [ʊ̃]. The development is summarized in Figure 2. (We can assume the diphthongal variant {-ei} represents an intermediate stage between {-i} and {-e}, although this is not crucial for the fundamental point.) The right-branching path of the tree traces the darkening in the quality of nasalized vowels. Loss of nasalization, represented by the left-branching paths, prevents further darkening.

Something like the reconstruction of events sketched in Figure 2 is pretty much taken for granted by North Germanic dialectologists (e.g. Sandøy 1996, 133), but mapping the individual innovations lends strong support to the reconstruction as well as brings much needed clarity to the confusing geographical picture. The areal distributions of each innovation are shown in Figure 3. Areas manifesting a given innovation are shown in black. The 'greater than' symbol (>) may be read 'at least as far as'.

As Figure 3 shows, Lowering and Retraction as far as [ɑ] affected almost all va-

[3] F2 can be made audible by whispering the vowels (Catford 1988). Produced in sequence, it should be possible to hear a drop in pitch as we proceed from [i] to [u].



(A) Lowering I (>ei) (B) Lowering II (>e) (C) Lowering III (>æ)

(D) Retraction (>ɑ) (E) Rounding (>ɔ) (F) Raising (>ʊ)

FIGURE 3: Areal distribution of vowel nasalization and darkening

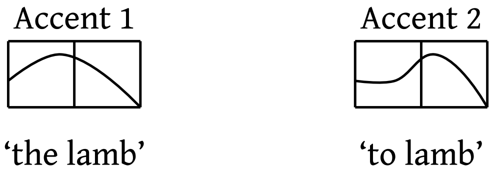
rieties of southern Norwegian (and, off-map, Swedish and northern Norwegian as well). However, western and mid Norway, the mountainous inland in particular, resisted these changes to various degrees. What we see is two kinds of conservatism at play. On the one hand, we see older high front vowel qualities being preserved inland. Most dialects show lowering at least as far as [ei] except two relic areas in grey in inland Agder and Telemark and Sognefjord (A), which preserve the [i] realization. Only slightly fewer dialects show lowering at least as far as [e] (B). Lowering as far as [æ] has left quite a large contiguous relic area in the mid Norwegian inland (C), and when we consider retraction as far as [ɑ], the relic area extends all the way down to the west coast (D). On the other hand, we see that darkening has proceeded further in some coastal areas. This innovation depends presumably on the nasalization of the vowel being preserved longer than elsewhere in the Scandinavian Peninsula. Thus in Rogaland and southern Hordaland, we find raising at least to [ɔ] (E), and in the northern part of Rogaland and southern part of Hordaland, the quality has been raised all the way to [u] (F).

[3] DOUBLE PEAKED ACCENT IN CENTRAL SCANDINAVIA

The previous section showed how linguistic theory may be brought to bear to clarify the geographical picture. In this section, we show how dialect geography can clarify the linguistic history.

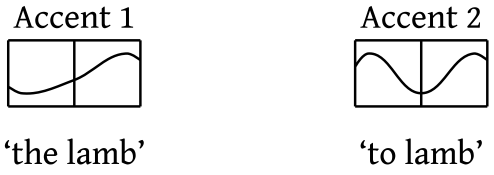
The North Germanic varieties spoken on the Scandinavian peninsula (Norwegian and Swedish) distinguish between two so-called word accents, Accent 1 and Accent 2. The tone accent contrast is exemplified in the stylized tone curves in (3) for the citation forms /ˊlame/ 'the lamb' and /ˇlame/ 'to lamb' in one urban variety of Nordland Norwegian (Bodø).[4] The boxes represent disyllables, the vertical line in the middle the syllable boundary.

(2) *Accent 1 vs. Accent 2 in Nordland Norwegian*



The way in which these accents are realized varies geographically. A shibboleth of what many non-Scandinavians take to be prototypically Swedish or Norwegian is a 'double-peaked' realization of Accent 2, illustrated for Oslo speech in (3).

(3) *Accent 1 vs. Accent 2 in Oslo Norwegian*



In a few Norwegian dialects spoken in the area of Trøndelag, both Accent 1 and Accent 2 have a double-peaked realization, the first peak occurring earlier in Accent 1. However many varieties have a single-peaked realization. The question which of these two realizations is the oldest has occupied scholars for decades. However, reconstructing accent invites special problems because it is not represented in the manuscripts. In the literature, both points of view have been defended. Elstad (1980), Lorentz (2005), and Hognestad (2006, 2007) argue that it is the single-peaked realization that represents the original state of affairs. In a series of papers over the last fifteen years or so, however, Tomas Riad has argued that it is the double-peaked realization that is the older of the two. Riad argues that the double-peaked realization may be reconstructed as far back as what is known as the Syncope Period of Proto-Nordic (esp. Riad 1998, 2003). The

[4] The diacritics /ˊ/ and /ˇ/ mark the primary stress as being associated with Accent 1 and Accent 2 respectively.

single-peaked realization, on Riad's view, is a later innovation. It can be hard to find compelling phonological or phonetic reasons for preferring either of these competing proposals over the other. What I propose to show here, is that, Riad's proposal may be challenged on geographical grounds independent of phonetic and phonological considerations. When we compare the areal distribution of the double-peaked realization with other known, approximately datable innovations, we in fact find a striking match. The picture that emerges is one where a large contiguous area of Central Scandinavia forms a relatively innovative block, with Trondheim as the main foundry of change. This also makes sense in the light of the archaeological and historical evidence.

[3.1] *The evolution of double peak*

Dialectal variation in the realization of the Accent 1–Accent 2 distinction was first mapped in a now classic study by Meyer (1937, 1954), who elicited disyllabic simplex words with declarative intonation for 100 North Germanic dialects, 93 of which were dialects of Swedish, 5 Norwegian, and 2 Danish. This prepared the way for a rich crop of further work, although until recently work on lexical tone has tended to be pursued along national lines. Significant work on the individual languages include Gårding & Lindblad (1973), Gårding (1975, 1977), Bruce (1977, 1983), Bruce & Gårding (1978), and Gårding et al. (1978) (Swedish); Fintoft & Mjaavatn (1980) (Norwegian).

Much work on North Germanic accent to date assumes the existence of lexical tones. Our starting point, however, is recent work by Morén (2007), who argues that the accent distinction is not tonal, but involves a difference in prosodic structure. The distinguishing feature of Accent 1 is here taken to be that the prosodic word contains a nested monosyllabic prosodic word.[5] Thus the Accent 2 word *lamme* 'to lamb' contains a single prosodic word $(\,\text{lamə}\,)_\omega$, while the corresponding Accent 1 word *lammet* 'the lamb' has a minimal prosodic word $(\,\text{lam}\,)_\omega$ contained within a higher-level (maximal) prosodic word containing both the root and the definite clitic *-et*: $(\,(\,\text{lam}\,)_\omega\ \text{ə}\,)_\omega$. Recursion at the level of the prosodic

[5] Stress in Standard Swedish falls by default on the penultimate syllable. Morén (2007) addresses a neglected correlation between accent and prosodic structure and stress, showing that exceptional finally as well as antepenultimately stressed words invariably have Accent 1. The problem of how to represent Accent 1 phonologically is therefore intimately, not to say essentially, connected with prosodic structure and the representation of lexical stress. Assuming diacritic marking is undesirable, encoding lexical stress must minimally entail representing the prosodic word node and its designated head, e.g. $/\overset{\omega}{\overset{|}{\text{du}}}\text{minu}/$ ˈ*domino*. The basis of the North Germanic accent distinction, interpreting a suggestion by Morén-Duolljá (*voce*), is that underlyingly specified prosodic words cannot acquire additional syllable nuclei due to some faithfulness requirement. Remaining nuclei that cannot be parsed into the underlying prosodic word must therefore be parsed into a higher level constituent, identified here as a recursion of the prosodic word. The lower (underlying) prosodic word is thus forced to surface as monosyllabic, and this structure has an effect on the intonational pattern. This account also echos a recent proposal by Lahiri et al. (2005) to the effect that Accent 1 is the marked member of the opposition.

word has also been argued by Itô & Mester (2006, 2008). The tonal component of both accents is purely intonational and, underlyingly at least, phonologically invariant across both accents. For purposes of this paper, I will take the difference in pitch pattern between Accent 1 and Accent 2 in any given dialect to be a matter of phonetic interpretation only: in Accent 2, the peak is timed to occur later relative to the beginning of the (minimal) prosodic word. This understanding also meshes with the finding that the length of the word correlates positively with the degree of peak delay in several languages. Longer words evince longer peak delays in English (Steele 1986; Silverman & Pierrehumbert 1990; Bruce 1990; House & Wichmann 1996), German (Grabe 1998), and Inis Oirr Irish (Dalton & Ní Chasaide 2003, 2007).

The null hypothesis is that the prosodic structure of Accents 1 and 2 is invariant across Peninsular North Germanic dialects; it is on the level of intonation that they vary. This makes the problem of reconstruction much more tractable. Before we grapple with the details of the reconstruction, however, let us briefly review the basics of intonational phonology. For two good recent introductions to this field, see Gussenhoven (2004) and Ladd (2008).

A basic distinction is generally drawn between pitch accents, which associate to stressed syllables, and boundary tones, which align to the edges of prosodic constituents, such as the intonation phrase.

Bye (2010) argues that cross-dialect accentual variation is the result of two kinds of phonetic enhancement and subsequent phonological reinterpretation of the output by new generations of speakers. Enhancement of the first kind involves delay of a high tone peak relative to the stressed syllable that is phonologically associated to the high tone. As Farrar & Nolan (1999) and Gussenhoven (2004) argue, delaying the peak makes it sound higher. This effect is apparently due to the way in which listeners exploit phonetic knowledge. Listeners tacitly know what pitch it is possible to achieve within a given interval. When the peak is delayed, listeners subconsciously add the extra pitch that it would have been possible to achieve within the onset-to-peak interval to the objective pitch. Peak delay is one of the strategies that speakers recruit in order to convey paralinguistic meanings deriving from what Gussenhoven calls the 'Effort Code', the tacit knowledge of the positive correlation between the size of the pitch excursion and the degree of effort required to achieve it. Although this correlation is at base physiological, speakers are nevertheless able to bring it under cognitive control and exploit it for communicative purposes. Once this occurs, the relation between pitch and the meaning it conveys, albeit still a scalar one, is to some extent conventionalized. Greater pitch excursion is universally interpreted on the affective level as indicating greater surprise, helpfulness or engagement; on the informational level, it signals greater urgency, and is frequently grammaticalized as a marker of focus. What we might call the 'exchange value' of pitch, that is, what degree of pitch excursion is necessary to signal a given level of engagement, will vary to some extent from one community to the next. In some communities, the use of a relatively wide pitch range may be semantically neutral, no more than a general characteristic of the speech of that particular community. In other speech communities where the pitch range is generally narrower, the same wide pitch span would be interpreted as semantically marked. Peak delay is a cost-effective way to signal these meanings because it results in higher perceived pitch with smaller excursions and less effort. Another strategy for making peaks sound higher without raising the objective pitch is by introducing a relatively low on-glide; valleys may similarly be enhanced by relatively high on-glides. Given the social value of signaling meanings such as engagement and helpfulness, speakers may be expected to exaggerate the apparent degree of pitch excursion beyond community expectations. After a while these expressive strategies undergo a kind of semantic bleaching and become entrenched as the neutral idiom in much the same way as happens with expressive vocabulary in general (e.g. swearwords, politeness formulae, and so on).

The phonetic enhancements just described provide the raw material for phonological reanalysis in the next generation. Figure 4 sketches the evolution of the double-peaked accent as proposed by Bye (2010). Each box shows the prosodic and tonal representations of the Accent 2 word *himmel* ‘sky’ and the Accent 1 word *segel* ‘sail’ at consecutive stages of development. The curve provides a visual representation of the phonetic realization. The left-to-right arrows mark changes in phonetic realization (the curve), i.e. peak delay and the introduction of enhancing high and low on-glides. The arrows going southwest mark phonological reanalyses, realignment of tones and the introduction of new tones. The most conservative varieties simply have a H pitch accent followed by a low boundary tone (L%) marking the right edge of the intonational phrase, giving a falling pattern over all. Once delay of the H tone peak occurs, it may be further enhanced by introducing a low on-glide, which is reanalyzed as the insertion of a phonological low tone. In Figure 4, this is shown in the change of an original H pitch accent into a bitonal $\widehat{\mathrm{LH}}$ complex. Since the low tone is a new target, it may become the object of enhancement itself, for example through the introduction of a relatively high on-glide. This is the phonetic origin of the double-peaked realization that is so characteristic of Central Scandinavia. The on-glide may also be phonologized as a high tone in a later generation, i.e. $\widehat{\mathrm{HLH}}$. Because the peak is timed to occur late in Accent 2 and early in Accent 1, this new initial H tone will only be audible in Accent 2. Here we set aside the question whether this entails a phonological difference or not.[6] This gives rise to tonal crowding and what is the linearly second high tone (a reflex of the original high tone) in the tritonal cluster is no longer accommodated within the stressed syllable. This creates an ambiguity for the learner: is the second high tone aligned to the stressed syllable (part of the pitch accent), or is it associated to the edge of the intonation phrase? These situations, where the learner lacks sufficient evidence to accept or reject a hypothesis about the structure of his language, create the conditions favorable for reanalysis to take place. In many dialects with the double-peaked realization, the second high tone has accordingly been reanalyzed as a H% boundary tone, giving rise to a cluster of boundary tones on the right edge. In the most progressive dialects, the original final L% boundary tone has been truncated. Finally, in the most advanced dialects, Accent 1 has also acquired a double peak with the introduction of a high on-glide.

[3.2] *The areal distribution of double peak*

If what was outlined in the previous section is the correct understanding of the evolution of the double-peaked realization of the accent, it should be possible to correlate each hypothesized innovation with a contiguous region on the map (Bàrtoli's areal norm (1-b)). The areas associated with later innovations should be nested within the areas associated with earlier ones deriving from the same centre of innovation (Bàrtoli's areal norm (1-c)).

Applying Bàrtoli's norms to double peak

Applying Bàrtoli's norms to the areal distributions of each innovation suggests that double-peaked realization is a Central Scandinavian innovation. By hypothesis, double-peaked accent arose in those varieties which had earlier shifted from early accent to delayed accent, which are shown on map (A) in Figure 5. This area properly includes the area in which double-peaked varieties are found, which is map (B).[7] A contiguous part of the double-peaked area, that includes eastern Norway and western Sweden, evinces truncation of the final low tone (C). Finally, in Trøndelag, the realization of Accent 1 is approaching that of Accent 2 by the addition of a high on-glide (D).

The double-peaked realization is thus largely found in one contiguous area (with a few exceptions explained below), whereas the single-peaked pronuncia-

[6] This issue commonly arises in discussions of truncation in the literature. Where the domain for the realization of some tonal contour is short, a tone at the edge of the domain may fail to be realized. It is not always a simple matter to decide whether to ascribe this effect to phonetic implementation or to a phonological deletion rule.

[7] The reader will notice that there is an outlier in the north of Norway which also has the double-peaked realization. Massive flooding in the southeast Norwegian inland in 1789 lead to the migration of large numbers of farming families in Østerdalen and Gudbrandsdalen, in southeastern Norway, to Bardu and Målselv.

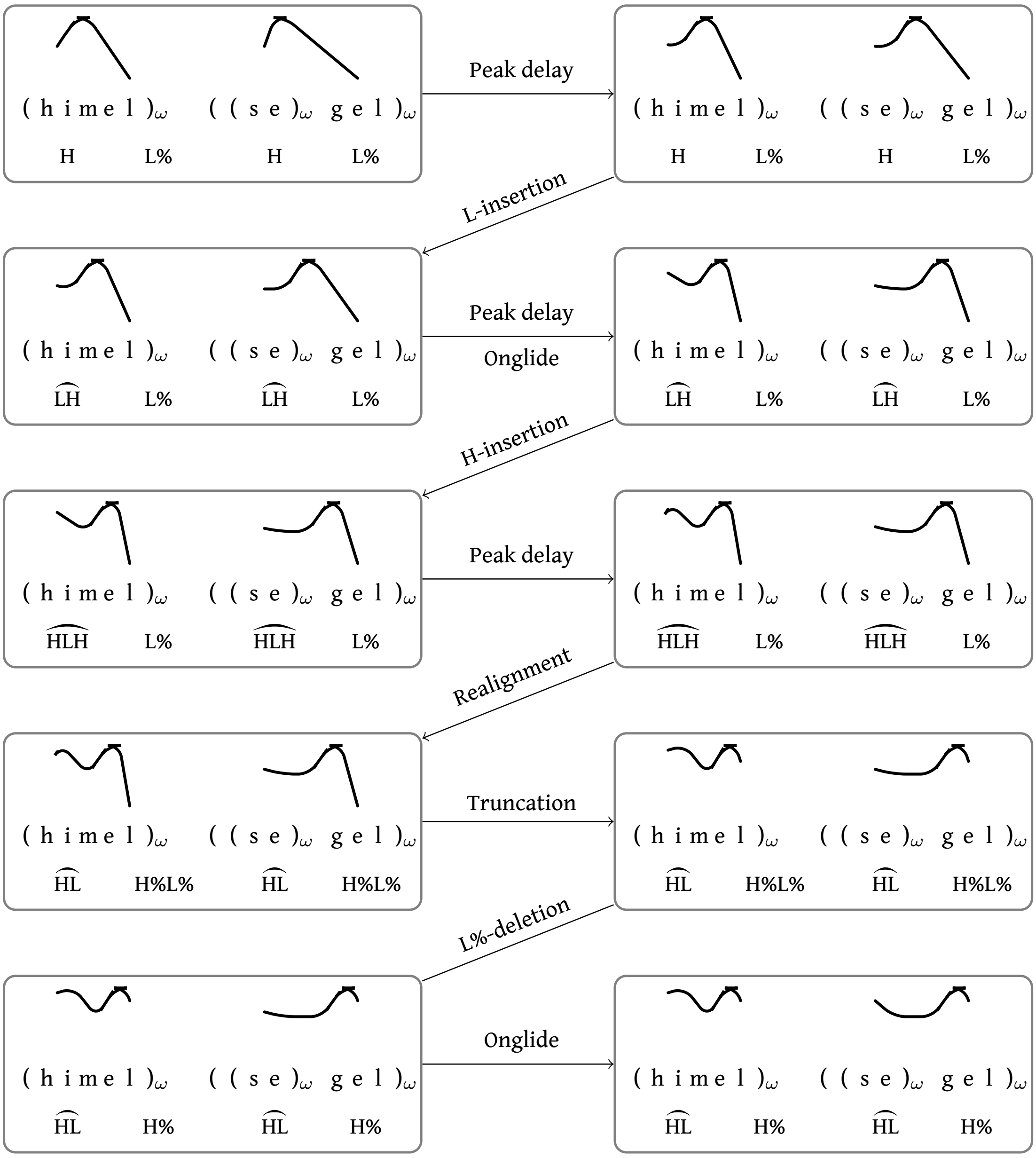
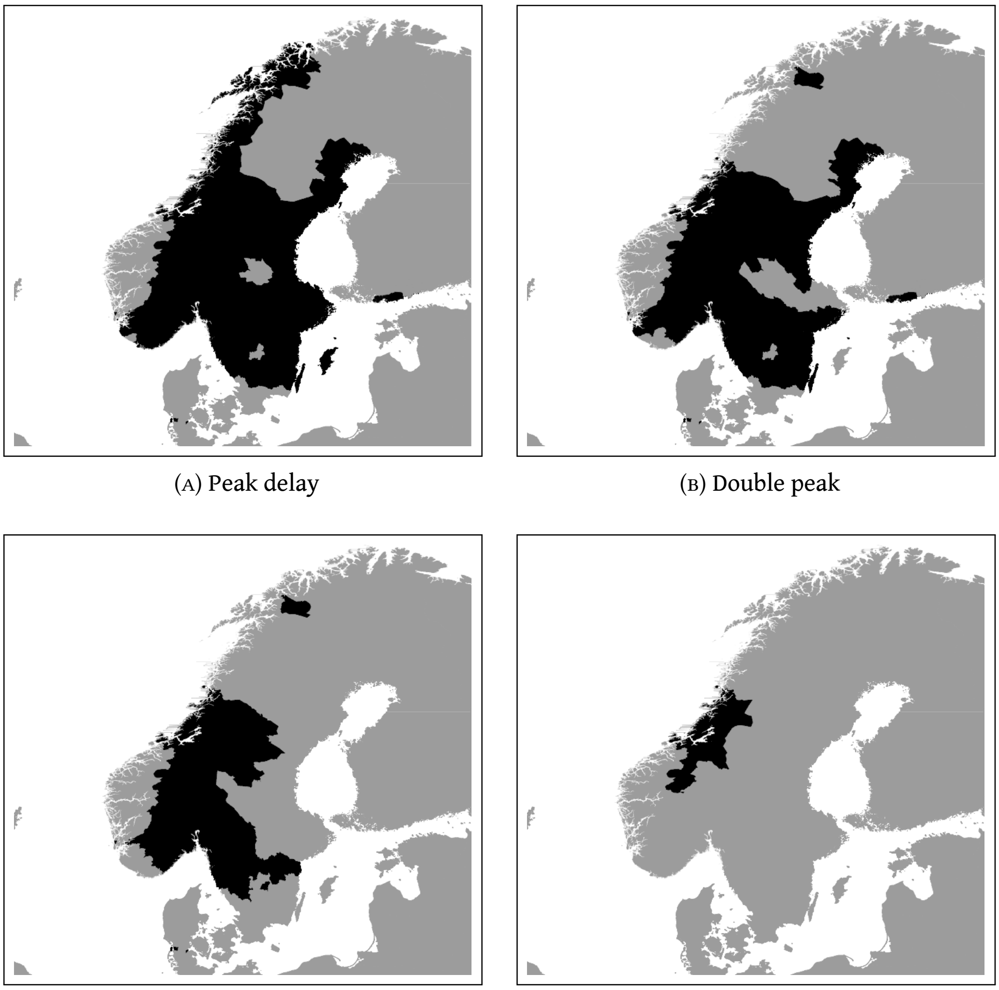


FIGURE 4: Evolution of double-peaked accent

tion is found in several separate areas. It is also found in isolated areas, such as Dalecarlia, peripheral areas such as Scania, West and North Norway, and its geographical extent is larger, reaching from Scania to Northern Norway. Given this, one can argue that it is unlikely that the single-peaked realization is the innovation, since it is unlikely that similar innovations start in separate areas. As an argument for the diachronic priority of single- over double-peaked Accent 2, it is far from watertight. For one thing, it is not unheard of for similar innovations



(A) Peak delay

(B) Double peak

(C) Truncation

(D) H on-glide in Accent 1

FIGURE 5: Central Scandinavian innovations in intonation

to arise in different places.[8] It is thus still possible that Riad's Hypothesis is correct. Indeed, Riad challenges the idea that the double-peaked realization could be the innovation on two additional grounds. First, it is possible to read the map a different way. There are several apparently outlying points on this map where double-peaked realizations of Accent 2 are found: Bardu and Målselv (northern

[8] One well-known case involves the phonemic split of Middle English /ʊ/ into /ʊ/ and /ʌ/. This occurred in two non-contiguous areas, Scotland and the Southeast of England. The split may nevertheless have been a single innovation that spread via social network connections between Glasgow, or Edinburgh, and London. The well-known glottal stop associated with vernacular forms of London English is another innovation that is reputed to have spread directly from Glasgow relatively recently (the earliest descriptions of Cockney lack glottal stop). See Andrésen (1968) and Fabricius (2000) for details.

Norway), Nyland (in Finland), and southern Denmark.[9] This may be interpreted to mean that it is the double-peaked realization that is the oldest. Second, Riad casts doubt on the notion that Central Scandinavia could have been a spreading zone, since it requires us to believe that the features were spread over putatively difficult terrain. This view is bound up with the fact that Riad places the development of accent very early — in the Proto-Nordic period, at a time when, he believes, "the land divides, the sea unites". We shall refute this view below.

Let us briefly comment on the apparent outliers. The inhabitants of Bardu and Målselv are largely the descendants of speakers of East Norwegian (which has a double-peaked realization of Accent 2) from Østerdal and Nordgudbrandsdal who migrated there in the 18th and 19th centuries. A number of varieties in southern Denmark are claimed to have double-peaked Accent 2. This claim requires much further investigation, however, since the available phonetic descriptions are sketchy and impressionistic. A historical connection between the double-peaked realizations of Accent 2 in southern Denmark and those in Central Scandinavia is in any event unlikely due to radical differences in the distributions of the accents in the two dialect groups. Kroman (1947) shows that, in South Funish, Accent 2 is restricted to disyllables whose root vowel was *short* in Common North Germanic. Disyllables whose root vowel was long have Accent 1. In Central Scandinavian on the other hand, all Common North Germanic disyllables evolved Accent 2. It is possible that the double-peaked realization of Accent 2 found in Nyland should be understood as part of a wider Central Scandinavian innovation. As we shall see below, shared innovations between East Swedish and Central Scandinavian lend some plausibility to this hypothesis. Other varieties of East Swedish (spoken in Finland and Estonia) have lost the lexical accent distinction, apparently quite recently as an effect of contact with Finnish (Ahlbäck 1971). Unfortunately, there is no data with regard to how the accents were realized in these varieties, so it is no longer possible to tell whether Nyland was part of a contiguous region of accentual innovation.

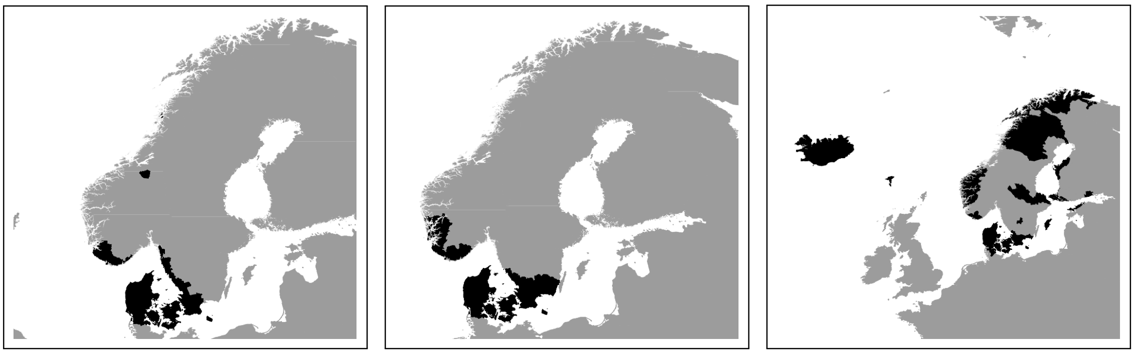
Double peak in geographical context

The relative age of features cannot be established on areal distributions alone. Where possible, they must be compared to those of known and datable innovations. This provides a fix on the time and place of origin of the putative innovation. Finally, this geographical picture must be related to the available ethnological, archaeological and historical evidence. This will be the topic of the final section.

Understood as an innovation, Single-peaked Accent fails to cluster with known

[9] Riad also mentions Stavanger as an outlier, but the maps of Fintoft & Mjaavatn (1980) show that there is a corridor from the inland to the west coast of varieties with a double-peaked realization of Accent 2. This realization is nonetheless new on the west coast. For details, see Hognestad (2006).

coastal innovations. The two most uncontroversial coastal innovations are the lenition of /p t k/ to [b̥ d̥ g̊] and the uvular realization of /r/. On the southern coast of Norway, for example, *mat* 'food' is pronounced [maːd̥] (standard: [mɑːt]). In Danish, [d̥] underwent further lenition to [ð̞]: [mɛð̞]. The uvular realization of /r/, which, as a broader European phenomenon, is found throughout much of the European continent as well. Its spread to the Danish capital in the late eighteenth century and, from there, to the southwest coast of Norway is a matter of historical record (Nielsen 1959).



(A) Lenition

(B) Uvular *r*

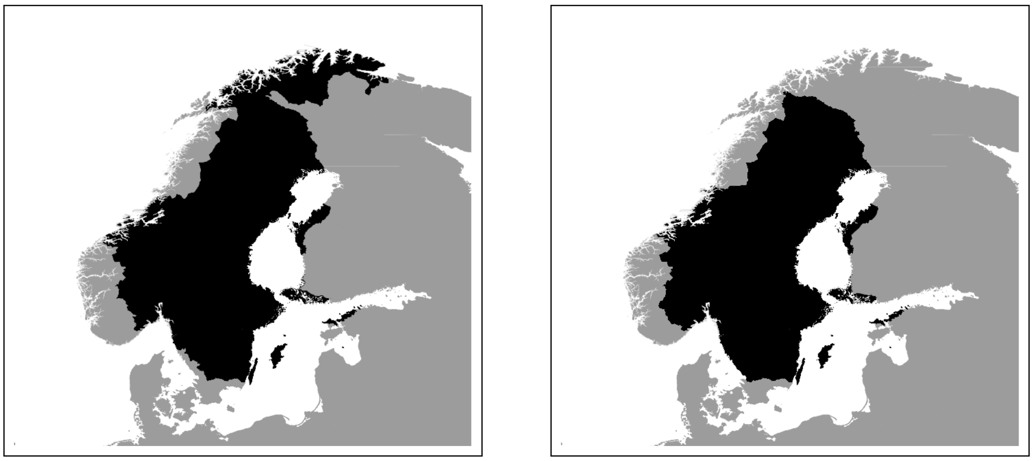
(C) Single peak

FIGURE 6: Two coastal innovations and single peak compared

The single-peaked realization is indeed also found in most of Denmark, Skåne, and the southwest of Norway, roughly the same areas where we find lenition and uvular *r*. However, it is also found in a contiguous region of Central Sweden, and northern Norway and Sweden. If we abstract away from the accent distinction, there is no reason not to include Finnish and Estonian varieties of Swedish, Iceland and the Faeroes, or for that matter most of Europe and beyond. The geographic evidence for a connection between single-peaked accent and innovations known to centre on Skagerrak is therefore weak.

Let us now turn to the clustering of the accentual innovations described in the two preceding sections, including the emergence of the double-peaked realization itself, with Central Scandinavian distribution. Two syntactic innovations are of broad Central Scandinavian provenience, shown in Figure 7. The first is the use of the expletive *det* in presentation sentences of the type *Det er kommet en båt* 'a boat has arrived' rather than *der* 'there', as in its geographical competitor *Der er kommen en båt*. Another Central Scandinavian feature is the use of *ha* 'have' in resultatives, e.g. *Hun har kommet hjem* 'she has come home' rather than the verb to be, as in *Hun er kommen hjem*.

Now let us turn to phonological features. One striking feature of the phonemic inventories of most Central Scandinavian varieties of North Germanic is the presence of retroflex consonants. Old Norse *l* became a retroflex flap [ɽ] in some en-

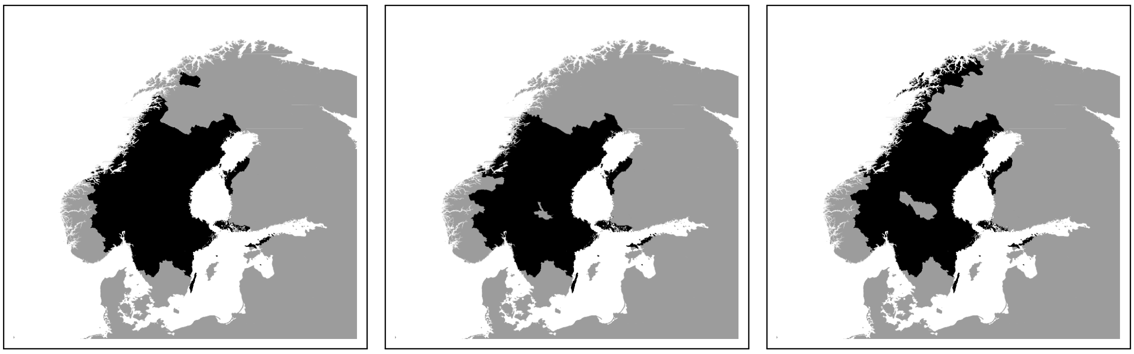


(A) Expletive *det*

(B) Resultative *ha*

FIGURE 7: Two syntactic innovations

vironments (Figure 8A). In a slightly smaller properly included area, Old Norse *rð* also became [ɽ] (Figure 8B). Another innovation was the development of retroflex consonants from clusters of /r/ followed by a coronal consonant /t d n l s/ (Figure 8C).



(A) Flap [ɽ] < *l*

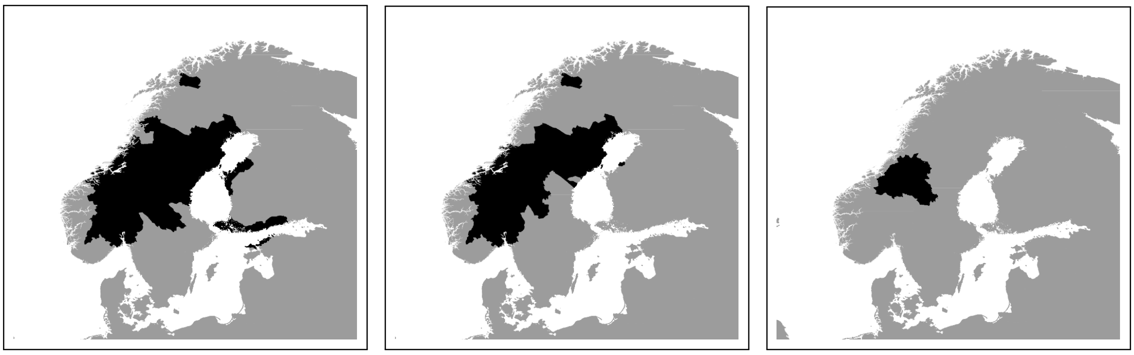
(B) Flap [ɽ] < *rð*

(C) Retroflexion

FIGURE 8: Retroflexion and the retroflex flap [ɽ]

Another important set of phonological innovations concerned the vowel system shown in Figure 9. Vowel balance refers to allophony in the desinential vowel. After a heavy root syllable $\bar{\sigma}$, as in a word like Old Norse *bíta* 'bite', the quality of the desinential vowel was reduced, e.g. $\bar{\sigma}$.Ci > $\bar{\sigma}$.Ce, $\bar{\sigma}$.Ca > $\bar{\sigma}$.Cɐ, $\bar{\sigma}$.Cu > $\bar{\sigma}$.Co. Following a light root syllable $\breve{\sigma}$, however, as in Old Norse *vita* 'to know', the quality of the desinential vowel was preserved, e.g. $\breve{\sigma}$.Ci, $\breve{\sigma}$.Ca, $\breve{\sigma}$.Cu. The areal distribution of these innovations is shown in Figure 9A. Related to this development in many dialects is metaphony of the root vowel in words with a light root syllable (e.g.

Bye 2008). All dialects with this feature minimally harmonize a low root vowel with a following /ɔ/, e.g. tala > tɔɽɔ 'speak', and (vacuously) sofa > sɔʋɔ 'sleep' (partial metaphony, Figure 9B). In a smaller area, metaphony has spread to all root vowels (full metaphony, Figure 9C); examples with desinential /ʉ/: ʋiku > ʋʉkʉ 'week', legu > lʉgʉ, hɔku > hʉkʉ, loku > lʉkʉ, furu > fʉrʉ 'pine'; examples with desinential /ɔ/: bita > bɔtɔ 'bite', skera > ʂɔrɔ 'cut', tala > tɔɽɔ 'speak', sofa > sɔʋɔ 'sleep', bruna > brɔnɔ 'thaw'.



(A) Vowel balance (B) Partial metaphony (C) Full metaphony

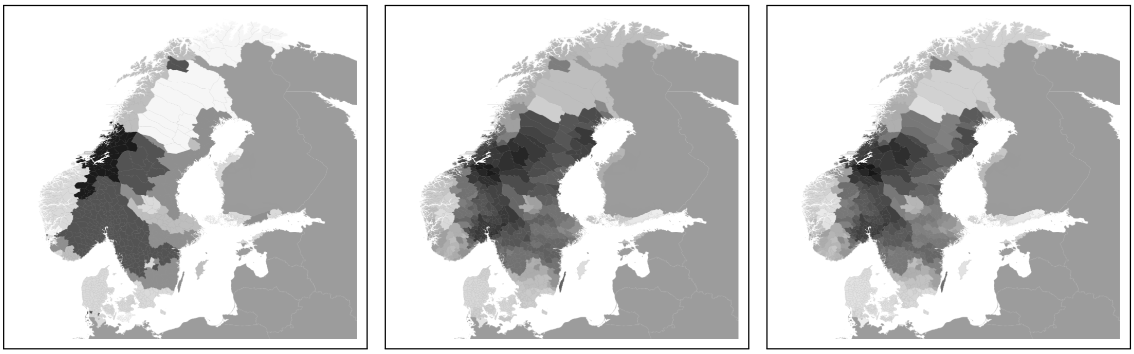
FIGURE 9: Vowel balance and metaphony

There is a good match between our hypothesized accentual innovations and other independently motivated, known and (reasonably) datable Central Scandinavian innovations. Putting all this together, it is possible to get a picture of where the geographical core of these innovations lies. In Figure 10, darker regions of the map are connected with more innovations, lighter regions with fewer. As we can see from 10A, it is Trøndelag that is the most innovative accentually, followed by southeastern Norway, and Jämtland and Götaland in Sweden. Considered against the non-accentual innovations (Figure 10B), Trøndelag is also the core. Figure 10C combines both accentual and non-accentual innovations into a single map. Southeastern Norway around Oslofjord and the Bothnian coast of Sweden are also dark. As can be seen, there is no evidence for the dictum that "the land divides, the sea unites".

Cultural, archaeological and historical context

It is fruitful to consider linguistic innovations in a broader archaeological and historical context.

Study of the maps of the previous section reveal that Trøndelag forms the hub of the Central Scandinavian 'province'. Full metaphony is associated with Trøndelag and Jämtland, and double-peaked realization of Accent 1 is associated with Trøndelag, Møre and Romsdal. Retroflexion and the flap also extends into Götaland in Sweden. It is therefore reasonable to suppose that Central Scandinavian



(A) Accentual innovations (B) Non accentual innovations (C) All innovations

FIGURE 10: Geographical core and periphery in Central Scandinavia

innovations in general can be traced back to the influence of Trøndelag. The primary axes along which these innovations spread can be ascertained by studying the maps for vowel balance (Figure 9A) and partial metaphony (Figure 9B). These maps suggest two primary axes of spread, one between Trøndelag and the Baltic coast of Central Sweden, and another between Trøndelag and Oslofjord. A secondary axis of spread may have existed between Oslofjord and Götaland. Below we shall review the archaeological and historical evidence for the importance of Trøndelag as a centre of influence and its connections with the rest of Scandinavia, and examine the nature of the contact.

It is clear that there has been a strong continuous connection between Trøndelag and the Baltic coast of Central Sweden, although the centre of influence on the Baltic coast has changed over the centuries. According to Elgvin (1961) there was an important trade route from the Møre and Trøndelag coast to the Baltic as early as the Stone Age. Archaeological evidence from the early Bronze Age (1–600 AD) also points to ancient and significant connections between Trøndelag and Middle Bothnia (Ångermanland and Medelpad) (Baudou 1986). Starting in the early Iron Age, however, Middle Bothnia was drawn into Mälardalen's sphere of influence. In the 7th and 8th centuries (the Merovingian period), there are a number of finds that attest to the presence Anglo-Saxon and Frankish influence in Trøndelag (e.g. the short sword known as the *scramasaxa*), but there are far stronger traces of influence from the Vendel culture of Uppland in Sweden (Marstrander 1956, 44) in the form of ornamented swords and spears, buckles with inlaid enamel, much of it associated with a new form of inhumed burials, generally boat graves. Taken in conjunction with what we know about the judicial organization of Trøndelag during this period and traditions about a connection found in the sagas, there is reason to believe there was a significant immigration from Svealand into Trøndelag in the Vendel era. In the Viking era, too, it is Trøndelag's *eastward* connections to the Baltic rather than its westward ties that really stand out. The large number of Arabian coins (a third of the ca. 400 found in Norway) attest to lively trading connections between Trøndelag and Mälardalen, in particular Birka, with its connections eastwards to Russia and the Black Sea. In contrast, the volume of finds in Trøndelag from western Europe is small.[10]

A route between Trøndelag and Oslo has been known since the Bronze Age and has always been the most important route in Norway (Steen 1942, 240). It is doubtful that this route itself was used for trade to any great extent, however. Its main purpose was administrative and ecclesiastical. Both Nidaros (now Trondheim) and Oslo were trading hubs in their respective areas, though, and there were also more southerly connections eastwards from East Norway into Sweden (Elgvin 1961). Those living in the inland valleys had to make journeys to Oslofjord, and the western fjords to stock up on salt and fish (Christiansen 1946, 56f.). Erixon (1933, 254) characterizes trade across the Norwegian-Swedish border as "lively and significant" (p. 254). There was a large market at Frösön in Jämtland where Trøndelag and Svealand came together to trade (Steen 1942, 82). As Erixon makes clear, trade was very largely import into Sweden of goods from Norway. For western Sweden, the nearest markets were located in Norway. Fish was transported from markets in Trondheimsfjord to the border, where Jämtlanders received the goods. The import business gave rise to a new breed of middlemen known as *färdmän* in Jämtland, Ångermanland and Härjedalen. Jämtlanders were also central in the market at Levanger in Trondheimsfjord where they sold iron tools and imported fish and horses (Hallan 1966).[11]

Horses were also an important in trade. During the winter, farmers from Upper and West Dalecarlia would transport horses, barrels of herring and dried cod from Trondheim or Röros via Lake Femund, Särna and Idre to the parishes around lake Silja. Erixon adds that this contact even extended across the Gulf of Bothnia into Finland. From the early mediaeval period the Bothnian Gulf became important for connecting western Sweden and the coast of Finland (Baudou 1987). Another shared characteristic is ornamental saddlery, which is found in East Norway from (Vestfold to Trøndelag), Dalarna, Härjedalen, Jämtland, Medelpad, Ångermanland, parts of Västmanland and northern Bohuslän.

Now that we have established that important routes existed connecting the Trøndelag–Baltic and Trøndelag–Oslo axes, it is worthwhile to consider the nature of the contact between the players (Christiansen 1946, 56f.). Traders travelled in large convoys, with 20 to 30 loads being common. These journeys took many

[10] This excludes a number of objects brought back from raids in Ireland, in which vikings from Trøndelag played a significant part. What is important here is enduring trade relationships.

[11] Jämtland, Härjedalen and Idre & Särna were part of Denmark-Norway until 1645 when they were ceded to the Swedish Crown as part of the terms of the Treaty of Brömsebro. After this the Swedish authorities began to monitor and keep records of Jämtlander trade for customs purposes. Hallan's evidence relates to this period but he stresses that there is every reason to believe that the Jämtlanders played an important role in the relations between Trøndelag and Svealand a long time before records began.

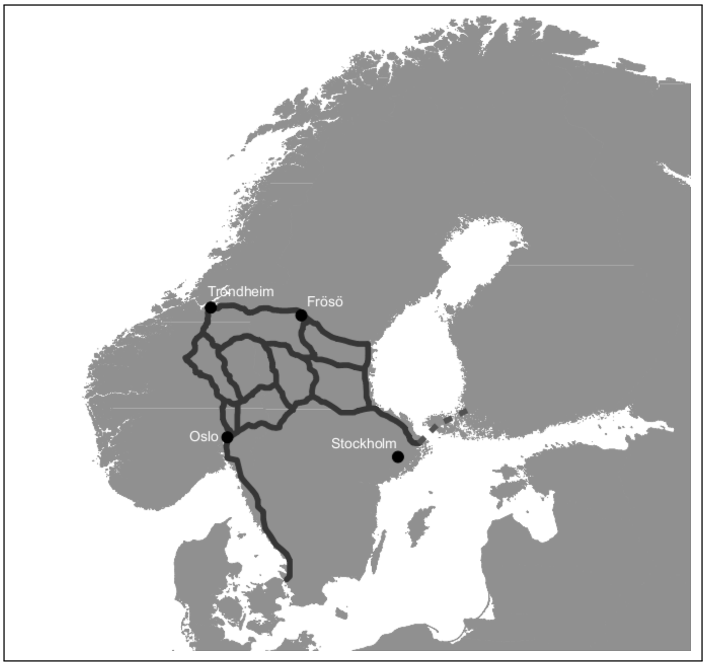


FIGURE 11: Central Scandinavian trade and pilgrimage routes

days. The journey from Gudbrandsdalen to Oslo, for example, took a fortnight with horse and packsaddle. Such journeys thus created conditions for forging lasting relationships with people outside one's local neighbourhood which would have encouraged accommodation of speech.

Also crucial for an understanding of the spreading of innovations are the routes of pilgrimage. These are shown in Figure 11 from information in Authén Blom (1961). The cult of Olav Haraldsson (St. Olav), who died at the Battle of Stiklestad in 1030 AD, was one of Europe's most important. After Rome itself, Olav's shrine in Kristkirken in Nidaros (now Trondheim) ranked in this respect alongside Santiago de Compostela and the shrine of Thomas à Beckett in Canterbury (Authén Blom 1961). Olav's Mass was celebrated every autumn. There were three main routes to Nidaros. The main route, which was used by pilgrims coming from the Continent (via Skåne and Konghelle) and from within Norway was the one through Oppland and Gudbrandsdalen described above. The route through Østerdalen was little used by Norwegians but was commonly used by Swedes (Steen 1942, 247). Interestingly, the cult of St. Olav was also widespread in Finland. Pilgrims from Finland would cross the Bothnian Gulf at Åland and proceed to Nidaros either following the more southerly route from Hälsingland to Härjedalen or the more northerly one over Jämtland and down into Verdalen. The latter, in addition to being the main arterial route between Sweden and Norway was also apparently especially popular with pilgrims because Olav Haraldsson himself took the same route on his return to Norway from exile in Novgorod. In sum, evidence from archaeology and history testify to the existence of robust networks throughout Central Scandinavia over which linguistic and other innovations could diffuse.

[4] CONCLUSIONS

Historical linguistics and dialect geography have much to gain from the use of geographical information systems to represent areal distributions of innovations. In this paper I have tried to show this with two examples from Peninsular North Germanic. In the first example, we saw how applying results from linguistic theory could clarify the geographic picture. Mapping innovations reveal a spatio-temporal structure that simply cannot be seen from the taxonomic map, which merely represents features. The second example showed how studying the areal distribution of a feature, and comparing its distribution to known innovations, can be used to support or refute a reconstruction.

REFERENCES

Ahlbäck, Olav. 1971. *Svenskan i Finland*, vol. 15, Skrifter utgivna av Nämnden för svensk språkvård. Stockholm: Norstedt, 2nd edn. First edition from 1956.

Andrésen, Bjørn S. 1968. *Pre-glottalization in English Standard Pronunciation*. Oslo: Norwegian Universities Press.

Authén Blom, Grethe. 1961. Pilegrimsveier. Norge. In Hødnebø (1956–1978), 306–310.

Bàrtoli, Matteo. 1925. *Introduzione alla Neolinguistica*. Genève: Olschki.

Bàrtoli, Matteo. 1945. *Saggi di linguistica spaziale*. Torino: Rosenberg & Sellier.

Bàrtoli, Matteo & Giulio Bertoni. 1925. *Breviario di neolinguistica*. Modena: Societa tipografica modenese.

Baudou, Evert. 1986. Ortnamn och nordliga kulturprovinser under järnålder och medeltid. In *Tre kulturer 3. Medlemsbok för johan nordlander-sällskapet*, 7–39. Universitetet i Umeå.

Baudou, Evert. 1987. Kontakter i Bottenviken från förhistorisk tid til medeltid. In Pekka Toivanen (ed.), *Bottnisk Kontakt III, 7–9.2.1986 i Jakobstad*, 7–11. Jakobstad: Jakobstads museum.

Beito, Olav T. 1986 [1970]. *Nynorsk grammatikk. Lyd- og ordlære*. Oslo: Det Norske Samlaget.

Bloomfield, Leonard. 1933. *Language*. Chicago: Chicago University Press.

Bonfante, Giuliano. 1947. The Neolinguistic Position. *Language* 23(4). 344–375.

Brøndum-Nielsen, Johannes. 1927. *Dialekter og dialektforskning.* Copenhagen: I. H. Schultz Forlag.

Bruce, Gösta. 1977. *Swedish Word Accents in Sentence Perspective*, vol. 12, Travaux de l'Institut de linguistique de Lund. Lund: Gleerup.

Bruce, Gösta. 1983. Accentuation and timing in Swedish. *Folia Linguistica* 17. 221–238.

Bruce, Gösta. 1990. Alignment and composition of tonal accents: comments on Silverman and Pierrehumbert's paper. In John Kingston & Mary Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, 107–114. Cambridge: Cambridge University Press.

Bruce, Gösta & Eva Gårding. 1978. A prosodic typology for Swedish dialects. In Eva Gårding, Gösta Bruce & Robert Bannert (eds.), *Nordic Prosody. Papers from a symposium*, vol. 13, Travaux de l'institut de linguistique de Lund, 219–228. Lund: Gleerup.

Bye, Patrik. 2008. Om oppkomsten og utviklingen av jamvekt og vokalbalanse i sentralskandinavisk. *Norsk lingvistisk tidsskrift* 26. 109–135.

Bye, Patrik. 2010. Tonal alignment and deep reanalysis in the evolution of Peninsular North Germanic tonal accent. Unpublished Ms., University of Tromsø/CASTL.

Catford, J. C. 1988. *A Practical Introduction to Phonetics.* Oxford: The Clarendon Press.

Christiansen, Hallfrid. 1946. *Norske dialekter I. Innføring i almen norsk fonologi og dialektologi.* Oslo: Tanum.

Christiansen, Hallfrid. 1969. *Norske målførekart.* Oslo.

Dalton, Martha & Ailbhe Ní Chasaide. 2003. Modelling intonation in three Irish dialects. In *Proceedings of the Fifteenth International Congress of Phonetic Sciences.*

Dalton, Martha & Ailbhe Ní Chasaide. 2007. Alignment and micro-dialect variation in Connaught Irish. In Tomas Riad & Carlos Gussenhoven (eds.), *Tones and Tunes. Volume 2: Experimental Studies in Word and Sentence Prosody.* Mouton de Gruyter.

Delvaux, Véronique, Thierry Metens & Alain Soquet. 2002. French nasal vowels: Acoustic and articulatory properties. In *Proceedings of the Seventh International Conference on Spoken Language Processing, Volume 1*, 53–56.

Elgvin, Johannes. 1961. Handelsveier. Norge. In Hødnebø (1956–1978), 168–170.

Elstad, Kåre. 1980. Some remarks on Scandinavian tonogenesis. *Nordlyd* 3. 62–77.

Erixon, Sigurd. 1933. Hur Norge och Sverige mötas. Studier rörande kulturgränser och kultursamband på Skandinaviska halvön. In John Frödin, Svend Aakjær, Sigurd Erixon & Alf Sommerfelt (eds.), *Bidrag til bondesamfundets historie II. Bosetning og kulturforbindelser*, 183–299. Oslo: Instituttet for Sammenlignende Kulturforskning / Aschehoug.

Fabricius, Anne H. 2000. *T-Glottaling between Stigma and Prestige: A Sociolinguistic Study of Modern RP*. Ph.D. thesis, Copenhagen Business School.

Farrar, Kimberly & Francis Nolan. 1999. Timing of F0 peaks and peak lag. In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, 961–964. San Francisco.

Fintoft, Knut & Per-Egil Mjaavatn. 1980. Tonelagskurver som målmerke. *Mål og Minne* 1980. 66–87.

Frings, Theodor. 1956. *Sprache und Geschichte*, vol. 16–18, Mitteldeutsche Studien. Halle: Niemeyer.

Gårding, Eva. 1975. Towards a prosodic typology for Swedish dialects. In K.-H. Dahlstedt (ed.), *The Nordic Languages and Modern Linguistics 2*, 466–474. Lund: Almqvist och Wiksell.

Gårding, Eva. 1977. *The Scandinavian Word Accents*, vol. 11, Travaux de l'Institut de linguistique de Lund. Lund: Gleerup.

Gårding, Eva, Gösta Bruce & Ursula Willstedt. 1978. Transitional forms and their position in a prosodic typology of Swedish dialects. In Eva Gårding, Gösta Bruce & Robert Bannert (eds.), *Nordic Prosody. Papers from a symposium*, vol. 13, Travaux de l'institut de linguistique de Lund, 197–206. Lund: Gleerup.

Gårding, Eva & Per Lindblad. 1973. Constancy and variation in Swedish word accent patterns. *Lund Working Papers* 7. 36–110.

Gordon, Eric V. 1981. *An Introduction to Old Norse*. Oxford: Clarendon Press, 2nd edn. Revised and updated by A. R. Taylor.

Grabe, Esther. 1998. *Comparative Intonational Phonology. English and German*. Ph.D. thesis, Nijmegen University.

Gussenhoven, Carlos. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Hägerstrand, Torsten. 1967. *Innovation Diffusion as a Spatial Process*. Chicago: University of Chicago Press. Translated from the original, *Innovationsförloppet ur korologisk synpunkt*, published in 1953 by C. W. K. Gleerup, Lund, Sweden. Translation and Postscript by Allan Pred.

Hallan, Nils. 1966. *Jemter på Levangsmarknaden i 1680-årene*, vol. 1, Skrifter utgivna av Jämtlands läns biblioteks vänner. Östersund: Wisénska.

Haugen, Einar. 1976. *The Scandinavian Languages. An Introduction to their History*. London: Faber and Faber.

Hødnebø, Finn (ed.). 1956–1978. *Kulturhistorisk leksikon for nordisk middelalder*. Oslo: Gyldendal.

Hognestad, Jan K. 2006. Tonal accents in Stavanger: From western towards eastern Norwegian prosody? In Gösta Bruce & Merle Horne (eds.), *Nordic Prosody. Proceedings of the IXth Conference, Lund 2004*, 107–116. Frankfurt am Main: Peter Lang.

Hognestad, Jan K. 2007. Tonelag i Flekkefjord bymål. *Norsk lingvistisk tidsskrift* 25. 55–88.

House, Jill & Anne Wichmann. 1996. Investigating peak timing in naturally-occurring speech: From segmental constraints to discourse structure. In Valerie Hazan, Stuart Rosen & Martyn Holland (eds.), *Speech, Hearing and Language: work in progress, Vol. 9*. London: University College London.

Itô, Junko & Armin Mester. 2006. Prosodic adjunction in Japanese compounds. In *Proceedings of The 4th Formal Approaches to Japanese Linguistics Conference (FAJL 4)*, vol. 55, MIT Working Papers in Linguistics, 97–112. Cambridge, MA: MIT Department of Linguistics and Philosophy.

Itô, Junko & Armin Mester. 2008. The onset of the prosodic word. In Steve Parker (ed.), *Phonological Argumentation: Essays on Evidence and Motivation*. London: Equinox.

Kranzmayer, Eberhard. 1956. *Historische Lautgeographie des gesamtbairischen Dialektraumes*. Graz: Böhlau.

Kroman, Erik. 1947. *Musikalsk akcent i dansk*. København: Einar Munksgaard.

Ladd, D. Robert. 2008. *Intonational Phonology*. Cambridge: Cambridge University Press, 2nd edn.

Lahiri, Aditi, Allison Wetterlin & Elisabet Jönsson-Steiner. 2005. Lexical specification of tone in North Germanic. *Nordic Journal of Linguistics* 28. 61–96.

Levander, Lars. 1909. *Älvdalsmålet i Dalarna. Ordböjning ock syntax.* Stockholm: Norstedt.

Lorentz, Ove. 2005. Tone shift and tone reversal in Scandinavian. Ms., University of Tromsø.

Marstrander, Sverre. 1956. Hovedlinjer i Trøndelags forhistorie. *Viking* 10. 1–69.

Meyer, E. A. 1937. *Die Intonation in Schwedischen. I: Die Sveamundarten*, vol. 10, Studies Scandinavicae Philologiae. Stockholm: Stockholms Universitet.

Meyer, E. A. 1954. *Die Intonation in Schwedischen. II: Die norrländischen Mundarten*, vol. 11, Studies Scandinavicae Philologiae. Stockholm: Stockholms Universitet.

Morén, Bruce. 2007. Central swedish pitch accent: A retro approach. Old World Conference in Phonology 4, Rhodes, 18–21 January 2007.

Nielsen, Nils Åge. 1959. Om bagtunge-r'ets opkomst i dansk. *Sprog og Kultur* 18. 58–64.

Petyt, K. Malcolm. 1980. *The Study of Dialect.* London: André Deutsch.

Riad, Tomas. 1998. The origin of Scandinavian tone accents. *Diachronica* 15. 63–98.

Riad, Tomas. 2003. Diachrony of the Scandinavian accent typology. In Paula Fikkert & Haike Jacobs (eds.), *Development in Prosodic Systems*, vol. 58, Studies in Generative Grammar, 91–144. Berlin: Mouton de Gruyter.

Rogers, Everett M. 2003. *Diffusion of Innovations.* New York: Simon and Schuster, 5th edn.

Sandøy, Helge. 1996. *Talemål.* Oslo: Novus.

Silverman, Kim & Janet Pierrehumbert. 1990. The timing of prenuclear high accents in English. In John Kingston & Mary Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, 72–106. Cambridge: Cambridge University Press.

Steele, Shirley A. 1986. Nuclear accent F0 peak location: Effects of rate, vowel, and number of following syllables. *Journal of the Acoustical Society of America* 80(S1). S51.

Steen, Sverre. 1942. *Ferd og fest. Reiseliv i norsk sagatid og middelalder.* Oslo: Aschehoug.

Trudgill, Peter. 1975. Linguistic geography and geographical linguistics. In Christopher Board, Richard J. Chorley, Peter Haggett & David R. Stoddart (eds.), *Progress in Geography. International Reviews of Current Research. Volume 7*, 227–252. Edward Arnold.

AUTHOR CONTACT INFORMATION

Patrik Bye
University of Tromsø
Department of Research and Development
N-9037 Tromsø
Norway
patrik.bye@uit.no

J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 31–43. (ISSN 1890-9639)
http://www.journals.uio.no/osla

LANGDOC: BIBLIOGRAPHIC INFRASTRUCTURE FOR LINGUISTIC TYPOLOGY

HARALD HAMMARSTRÖM AND SEBASTIAN NORDHOFF
Nijmegen and Leipzig

ABSTRACT

The present paper describes the ongoing project **LangDoc** to make a bibliography website for linguistic typology, with a near-complete database of references to documents that contain descriptive data on the languages of the world. This is intended to provide typologists with a more precise and comprehensive way to search for information on languages, and for the specific kind information that they are interested in. The annotation scheme devised is a trade-off between annotation effort and search desiderata. The end goal is a website with browse, search, update, new items subscription and download facilities, which can hopefully be enriched by spontaneous collaborative efforts.

[1] INTRODUCTION

Language Typology is the subfield of linguistics concerned with *the systematic study of the unity and variation of the languages of the world*. Like many disciplines, there are various infrastructural needs which are not yet in place. A central such need is as follows. Typically, the material for study for a typologist is a document with descriptive information on a language. With some 7 000 languages in the world (Lewis 2009), the number of relevant such documents grows far beyond the capacity of individual typologists. At present, single individuals have to manage micro-collections of references for their own use, which means not only gathering and re-typing them but also performing very time-consuming searches. The present paper describes a project **LangDoc** aimed at eradicating this enormous duplication of work, by providing a free and (if not complete) extensive collection of bibliographical references[1] available for download, search, subscription etc via a website.

In essence, the goal of LangDoc is as follows:

[1] Many of the actual documents that the references point to are difficult to access, tucked away only here and there in libraries across the world. Arguably, there is a similar superfluous duplication of work involved in accessing them. However, the present paper does not address this matter, which appears to be vastly more complicated than collecting only the references.

- Delineate a class of bibliographical references, namely those to descriptive materials
- Annotate them with focus (what language, family, etc.) and with type (wordlist, phonology, grammar etc.) such that
  - basic search criteria are met
  - the identity- and type-annotation has good automatization prospects
- Provide an updateable website interface

We will first define the scope of the proposed collection of references, and discuss some existing databases. Next, we will address issues of annotation and search desiderata. Finally, we will touch issues of update management, community contribution and crediting.

[2] THE SCOPE OF THE BIBLIOGRAPHICAL DATABASE

[2.1] *Desired Scope*

At present, a bibliography of all relevant research articles, e.g., 'all articles ever written in linguistics' or even 'all articles relevant for typology', however useful, seems much too large to be feasible. However, a bibliography of *descriptive* materials of the languages of the world is a fairly well-delineable class. For short, a bibliographic reference to a publication with descriptive/documentational data and/or metadata (number of speakers, location etc) will be called a **BDP**. The class of BDPs, as opposed to a mix of general linguistics articles, is of salient usefulness for a typologist. Furthermore, albeit with some work, it appears to be within scope to achieve a (near-)complete such database the following sense:

A) For every language, include the most extensive piece(s) of documentation, *and*

B) Beyond that, include "as much as possible"

This policy implies that

- for a small language with only a wordlist to its documentation, that BDP should be included
- for a bigger language with countless articles/books, a major dictionary/ grammar/text collection should be included, but not necessarily every single BDP ever written about the language (but, of course, any amount of these are also welcome)

[2.2] *Collecting References*

Language documentation and description is, and has been, an extremely decentralized activity. For well over two centuries, there has been intensive collection of data on the languages of the world by missionaries, anthropologists, travellers, naturalists, amateurs, colonial officials, and not least linguists. For natural reasons, all these people, including the linguists, hail from all parts of the world and call from maximally disconnected research environments. As a result, finding and tracking references to descriptive materials is not a straightforward task.

Traditionally, bibliographies would be curated by individual researchers, often experts on some area or language family, who happened to take on the matter after decades of collection, and then published in book form. These, when available and recent, are excellent guides, but do not cover the entire world (unless accumulated – see below), which is usually the frame of interest of the typologist. There are also a few bibliographies which have world-wide scope, but which are imperfect to the needs of the typologist in one or the other way. For example, the Ethnologue website[2] by SIL International lists references, but almost all of them are to works by SIL affiliated authors – a significant but small subset of the entire author space – and systematically excludes languages that went extinct before 1950, even if they are well-documented. The Linguistic Bibliography Online website[3] systematically fails to include MA/PhD theses and items from minor countries, and requires a subscription fee. The Worldcat catalogue[4] also fails to include many MA/PhD theses and other items for minor countries, and has no way of singling out linguistically relevant publications. Though some entries in Worldcat have annotation, overall, this is so unsystematic that it is of little use for finding BDPs on, e.g., a small Papuan language. Google, Google Scholar, and Google Books are, of course, resources with enormous coverage, but for browsing or zooming in on a specific language or area, it is difficult to come up with high-precision searches.

Now, given how decentralized language description is, one may doubt why it should even be possible to build a bibliographical database that meets high standards of completeness and precision. Who knows of all the obscure BDPs? We submit that experts of countries/language families/areas do tend to know the BDPs, obscure and non-obscure, of their respective field of interest. These experts write overviews and handbooks on a regular basis. For example, one type of overview with BDPs is a traditional printed book bibliography, such as:

> Newman, Paul. (1996) *Hausa and the Chadic Language Family: A Bibliography*.

[2] http://www.ethnologue.com accessed 1 Jan 2010. The printed edition in book form does not have all the references that the website has.

[3] http://www.blonline.nl/public/ accessed 1 Jan 2010. Printed editions in book form appear annually.

[4] http://www.worldcat.org accessed 1 Jan 2010.

Köln: Köppe [African Linguistic Bibliographies 6].

Another type of overview is a descriptive overview, i.e., an overview of what languages there are and a little about their nature in a certain area, such as:

> Laycock, Donald C. (1968) *Languages of the Lumi Subdistrict.* Oceanic Linguistics VII(1):36-66.

Further, perhaps the most common kind of overview with bibliographical references to the languages covered is a historical-comparative work, such as:

> Adam, Lucien. (1893) *Matériaux pour servir à l'établissement d'une grammaire comparée des dialectes de la famille Caribe.* Paris: J. Maisonneuve [Bibliothèque Linguistique Américaine XVII].

In addition, there are sociolinguistically oriented overviews, such as:

> Shearer, Walter and Sun Hongkai 2002 *Speakers of the Non-Han Languages and Dialects of China*, Lewiston, NY: Edwin Mellen Press [Chinese Studies 20]

and so on. Thus, going through all such overviews and handbooks collecting the references, is a systematic procedure for attaining a satisfactory world-wide bibliographical database. However, this only holds if there exist (recent) experts covering the whole world and that all their handbooks and overviews can be enumerated, since they, too, are of the same decentralized nature as the descriptive works on the languages themselves. The difference is that there are much fewer experts, areas, families and countries than there are languages, so the matter is more manageable. Nevertheless, the absolute number of overviews exceeds 5 000, according to our own collections so far.

[2.3] *Some Existing Resources*

Related to the above questions of how to collect and what to collect, significant headways have already been made in the actual work of doing the collection. Table 1 lists some existing resources of special interest to the present project.

All the resources of Table 1 are updated regularly, wherefore we report the time the information was collected. The Electronic Bibliography of African Languages and Linguistics (EBALL)[5] by Jouni Filip Maho, the Diccionario Etnologüístico y Guía Bibliográfica de los Pueblos Indígenas Sudamericanos (here abbreviated DEPIS)[6] by Alain Fabre, World Grammar Bibliography (WGB) by Harald Hammarström are bibliographies collected by single dedicated individuals following

[5] See http://goto.glocalnet.net/maho/eball.html, accessed 1 Jan 2010. WEB-BALL by Guillaume Ségérer is an online query interface that is based on an independently updated earlier version of EBALL (with ca 50% of the entries of the 2009 version) available at http://sumale.vjf.cnrs.fr/Biblio/index.html accessed 1 Jan 2010.

[6] See http://butler.cc.tut.fi/~fabre/BookInternetVersio/Alkusivu.html accessed 1 Jan 2010.

	# Refs	Contents	Area	Coverage	Annotation		Date
EBALL	60 164	Everything	Africa	Full	100%	L & T	Sep 2009
DEPIS	30 176	Everything	S America	Full	100%	L	Sep 2009
WGB	15 103	DD	World	85%?	100%	T	Dec 2009
MPIEVA	13 966	Everything	World	?	62-93%	L & T	Sep 2009
WALS	5 633	Mainly DD	World	?	99%	L	Aug 2005
SIL	18 464	Mainly DD & VP	World	70%?	100%	L & T	Sep 2009
SILPNG	13 110	Mainly DD & VP	Papua	Full	100%	L & T	Sep 2004

TABLE 1: Some existing bibliographical resources and their size, contents, annotation and the time the information was culled. Abbrevations are L = Language, T = Type, DD = Descriptive Data, VP = Vernacular Publications.

more or less the methodology outlined above; to go through all overviews. While EBALL and DEPIS strive to include everything, not just BDPs, on the respective languages, including all references to work done on relatively well-studied languages (such as Aymara or Hausa) and including non-descriptive work where the language in question is brought up (for example, in a discussion of the merits of a linguistic theory), WGB only strives to include the best descriptive work(s) on every language. This is the reason WGB has worldwide scope but is much smaller than the respective area-specialist bibliographies. MPIEVA is the online queryable library catalog of the Max Planck Institute for Evolutionary Anthropology[7]. In contrast to many other libraries, there is a dedication to collect descriptive data on the languages of the world, and most of the entries are annotated with ISO 639-3 codes, which makes it relatively simple to extract the part of the catalogue which refers to descriptive works. The WALS is a landmark multi-person typological project whose bibliography is ISO 639-3 annotated and available on the web[8]. The SIL Bibliography is the bibliography[9] of the missionary/linguist organization SIL International whose members have worked on a significant part of the world's lesser described languages. SILPNG (Akerson & Moeckel 1992; Linden 2003; Feldpausch 2005a,b) is a paper bibliography of the Papua New Guinea branch of SIL, where a significant part of the world's lesser described languages are found. SIL is a decentralized organization, and not all SILPNG references are included in the SIL Bibliography.

Access and license matters to the above collections are not yet clear, but it is likely that all of them can be used for benevolent purposes.

[7] http://www.eva.mpg.de/english/library.htm accessed 1 Jan 2010.

[8] http://wals.info/refdb/search accessed 1 Jan 2010

[9] http://www.ethnologue.com/bibliography.asp accessed 1 Jan 2010.

[3] ANNOTATION AND SEARCH DESIDERATA

[3.1] *Baseline Functionality*

Essentially, the typologist is looking for a BDP either from the language-side or from the document-contents-side (or a combination). Searching from the language side is typically to get whatever references are associated with a particular language, or associated with the language(s) that have some property such as 'belonging to family X' or 'endangered'. From the document-contents-side, the typologist may be looking for kinds of content of the document, such as 'contains wordlist', 'contains a section on adjectives' or 'contains interlinear glossed text'.

From the searcher's viewpoint, the more and the more detailed content-annotation the better, but from the annotators viewpoint, more and more detailed annotation is more and more work, unless the annotation can be (semi-)automatized. In general, we only have access to the text of the bibliographical reference itself (author, title, year etc.), not the actual document it refers to. Therefore, inferences depending on page counts or words that tend to occur in the title are possible, e.g., the name of the language(s) being treated often appears in the title (see below), but we cannot tell, e.g., whether there is a chapter/section on 'adjectives' or whether numerals are included in a wordlist.

Based on experience, the authors propose the following annotation scheme as a compromize between search desiderata, annotation work and (semi-)automatizabillity.

Identity: The language(s) the BDP treats. As a baseline, we suggest `ISO 639-3`[10] codes should be used as the identity registry. `ISO 639-3` codes are preferable as a baseline since linguists are used to them and they have good automatization properties. Furthermore, there already exists a database from which location, speaker number, genealogical classification etc. can be retrieved from `ISO 639-3` codes.

Other identity schemes, notably the doculect-languoid scheme (Cysouw & Good 2007; Good & Hendryx-Parker 2006) are more dynamic, and will in the end supersede the special status of the level of a maximal set of mutually intelligible varieties, which is the backbone of the `ISO 639-3` division (Lewis 2009, 7-18). For this reason, we also foresee a complementary, open-ended, identity annotation scheme which allows arbitrary (groups of) varieties on the sub-language level.

Type: The type/content of the document the BDP refers to. As a midway between our impression of typologists search desiderata, already existing annotation (e.g., from library catalogues) and (semi-)automatizability, we propose the following relatively uncontroversial hierarchy:

[10] See `http://www.sil.org/iso639-3/default.asp` accessed 1 Jan 2010

- (full-length) descriptive grammar
- grammar sketch
- dictionary
- description of some element of grammar (i.e., noun class system, verb morphology etc)
- phonological description
- text (collection)
- wordlist
- document with meta-information about the language (i.e., where spoken, (non-)intelligibility to other languages etc.)

We wish to stress the importance of partial automatizability of BDP annotation, which is some kind of guarantee that the endeavor will actually lead to a finished product and that updates are not very expensive.

As an example of how partial automatization of BDPs may work, we walk through an experiment described in Hammarström (2008) on how `ISO 639-3` language identity codes may be extracted from the title line of a BDP.

More formally, the problem may be cast as follows:

Given: A database of the world's languages (consisting minimally of <unique-id, language-name>-pairs)

Input: A bibliographical reference to a work with descriptive language data (= a BDP) of (at least one of) the language in the database

Desired output: The identification of which language(s) is described in the bibliographical reference

Unfortunately, the problem is not simply a clean database lookup! For example, a BDP might look as follows:

> Dammann, Ernst. (1957) *Studien zum Kwangali: Grammatik, Texte, Glossar*. Hamburg: Cram, de Gruyter & Co [Abhandlungen aus dem Gebiet der Auslandskunde / Reihe B, Völkerkunde, Kulturgeschichte und Sprachen 35].

This reference happens to be written in German. In general, the metalanguage could be any language (ca. 30 actually occur). The reference happens to describe a Namibian-Angolan language called Kwangali, `ISO 639-3` *kwn* and the task is to automatically infer this using a database of the world's languages and/or databases of other annotated bibliographical entries, but without humanly tuned thresholds. In the `ISO 639-3` database, each language has a three letter id, a canonical name and a set of variant and/or dialect names, for example

$Words(e_t)$	$LN(Words(e_t))$	$Words(e_t)$	$LN(Words(e_t))$
étude	$\{\}$	cameroun	$\{\}$
du	$\{dux\}$	du	$\{dux\}$
samba	$\{ndi, ccg, smx\}$	nord	$\{\}$
leko	$\{ndi, lse, lec\}$	famille	$\{\}$
parler	$\{\}$	adamawa	$\{\}$
d'allani	$\{\}$		

TABLE 2: For an example entry e_t, we show how many ISO 639-3 identities are associated with each word in the title of the entry.

Canonical name: Kwangali
ISO 639-3: kwn
Alternative names: {Kwangali, Shisambyu, Cuangar, Sambio, Kwangari, Kwangare, Sambyu, Sikwangali, Sambiu, Rukwangali}.

The languages and language name database consists of 7 299 languages, 42 768 language name tokens, 39 419 unique name strings. It is not yet well-understood how "complete" this language name database is, but as a rough indication we manually checked 100 randomly chosen bibliographical entries, whose titles contained a total of 104 language names. 43 of these names (41.3%) existed in the database as written, and 66 (63.5%) existed in the database, if one allows for spelling variation.

The size of the language name database is both a blessing and a burden. It may first seem as simple as looking up every word in the title of a BDP and pick the language whose name matches at least one word. Unfortunately, such a procedure only gets around 20% accuracy. To see why, consider the following example BDP:

> Fabre, Anne Gwenaïélle. (2002) *Étude du Samba Leko, parler d'Allani (Cameroun du Nord, Famille Adamawa).* PhD Thesis, Université de Paris III – Sorbonne Nouvelle.

The ISO 639-3 codes whose language name matches at least one word in the title is shown in Table 2. It so happens that such a common strings of letters as *du* happens to be a language name! The correct classification is this case is only $\{ndi\}$.

Clearly, we cannot guess blindly which word(s) in the title indicate the target language. But we can exploit some domain specific properties:

- A title of a publication in language description typically contains

  (i) One or few words with very precise information on the target language(s), namely the name of the language(s)

	foe	pole	huli	papua	guinea	comparativen	new	study	languages	and	a	the	of
$WC(w)$	1	2	3	57	106	110	145	176	418	1001	1101	1169	1482
Rel.Inc.	1.0	2.0	1.5	19.0	1.86	1.04	1.32	1.21	2.38	2.39	1.10	1.06	1.27

TABLE 3: The values of $WC(w)$ for w taken from an example entry (mid row). The bottom row shows the *relative increase* of the sequence of values in the mid-row, i.e., each value divided by the previous value (with the first set to 1.0).

(ii) A number of words which recur throughout many titles, such as 'a', 'grammar', etc.

- Most of the languages of the world are poorly described, there are only a few, if any, publications with original descriptive data.

Thus a more clever way is to divide the words in the title into two groups, informative and non-informative, and only use the informative ones for lookup. How can we measure the informativeness of a word w? Let $WC(w)$ = the number of distinct codes associated with w in the training data (set of already annotated BDPs) or Ethnologue database. Then for each word w, we get a value of informativeness. The question remains, at which point (above which value?) of informativeness do we get a near-unique language name rather than a relatively ubiquitous non-informative word? Luckily, we are assuming that there are only those two kinds of words, and that at least one near-unique language will appear. Thus, if we cluster the values into two clusters, the two categories are likely to emerge nicely. The simplest kind of clustering of scalar values into two clusters is to sort the values and put the border where the relative increase is the highest. The following example illustrates the method:

> W. M. Rule 1977 *A Comparative Study of the Foe, Huli and Pole Languages of Papua New Guinea*, University of Sydney, Australia [Oceania Linguistic Monographs 20]

Table 3 shows the title words and their associated number of codes (sorted in ascending order).

The highest relative increase is 19.0 between Huli and Papua. Thus, Foe, Pole and Huli are deemed near-unique and the rest non-informative. In this example, the three near-unique identifiers are correctly singled out.

The above method achieves about 70% accuracy, which can be slightly improved by allowing for spelling variants and disambiguation schemes (for details see Hammarström 2008).

So far we have not experimented with type-annotation, but impressionistically a similar level of accuracy seems achievable.

[3.2] *Extended Functionality*

Slightly more challenging than browsing for document properties is the browsing of language family trees. Depending on the scope of the research question, speech varieties smaller or bigger than the traditional 'language' are of interest. For instance, dialectologists will find it useful to narrow down their searches to the dialects of Croatian spoken in Italy instead of stopping at the language level of 'Croatian' `ISO 639-3 hrv` and be provided with information about Standard Croatian and other irrelevant dialects. On the other hand, comparatists will find it useful to have a node of all Scandinavian Northern Germanic languages together instead of having to collect the references for each language separately (`ISO 639-3 swe`, `ISO 639-3 nor`, `ISO 639-3 dan`, etc). This is even more relevant for less well-known language families and large-scale typology, where queries like "Give me a reference to every full description of a Nilotic language" are perfectly normal. It is therefore interesting to go beyond the flat list provided by `ISO 639-3` and add information about genetic nodes above and below the level of language as defined by the ISO-codes.

Existing genetic linguistic classifications can be exploited for this purpose. The multitree-project[11] contains a number of different linguistic classifications of the languages of the world in XML-format. Among these are so-called 'composite trees', which combine classifications of one family by different authors, diverging in scope and detail, into a much larger tree. These composite trees contain information about dialects as well as overarching large family classifications on a continental scale. A language typologist can select a node on the tree which corresponds to the scope of his or her study (dialect, language, language family, or any level in between). This node can then be used in database queries, together with the BDP properties mentioned above. A query on a node will return all documents which are attached to the node itself or any of its daughter nodes.

A major problem is that the assignment of BDPs to arbitrary nodes is more difficult to automatize than the assignment of BDPs to the standardized set of 7589 ISO-language names. For the time being we aim at attaching all BDPs to nodes which have an ISO-code as a start. Chosen users will be granted the right to reassign BDPs to other nodes interactively in a browser interface. Most typically, this will mean assigning a particular BDP to a subvariety below the node with the ISO-code, e.g.

[11] `http://linguistlist.org/multitree/` accessed 1 Jan 2010.

> Sammartino, Antonio. (2004) *Grammatica della lingua Croato-Molisana.* Zagreb: Fondazione "Agostina Piccoli".

would be reassigned from the node `<node name="Croatian" iso639-3="hrv">` to `<node name="Molise Croatian" iso639-3="">`. This graphical user interface will also allow users to add new BDPs and to assign them to the relevant nodes, assuring that the project will go with the times.

[4] ORGANIZATION AND MANAGEMENT

As already declared, the goal of LangDoc is a website with a comprehensive and annotated BDP bibliography with functionality such as browsing, searching, updating, new items subscription and downloading. BDPs have a well-defined structure and there are no interesting technical aspects of providing a web-interface to them.

At present, a functioning such website is not far away. However, it is useful to also consider how to best keep it updated, and how to make it a functioning collaborative resource. To encourage the submitting of additions/corrections by the public, and to give credit where credit is due, the information on who submitted the entry should be saved and displayed. Another option is to allow major resources to be "published" under the website's umbrella, with a clear identity surrounding it. The advantage of putting it under the umbrella would be that it is integrated in tools and search scopes of the overarching website.

[5] CONCLUSION

The present paper describes LangDoc, a project to make a bibliography website for linguistic typology, with a near-complete database of references to documents that contain descriptive data on the languages of the world. This provides typologists with a more precise and comprehensive way to search for information on languages, and for the specific kind information that they are interested in. The annotation scheme devised is a trade-off between annotation effort and search desiderata. In addition to saving time, such a database also has other uses. For example, there are so far unanswered questions about exactly how many and which languages of the world have been described, which have not, and which have partial descriptions. Another use has to do with the growing uneasiness of typologists towards the notion of language as a maximal set of mutually intelligible varieties. The typologist may also be interested in sub-language-level varieties and contrast between them, and may therefore want to build a catalogue of varieties (rather than languages). Such a catalogue of varieties is naturally based on the target documents of BDPs, and defining a variety reduces to saying which BDPs fall within it.

REFERENCES

Akerson, P. & B. E. R. Moeckel. 1992. Bibliography of the Summer Institute of Linguistics Papua New Guinea Branch 1956-1990. In *Summer Institute of Linguistics*. Ukarumpa, Eastern Highlands Province, Papua New Guinea.

Cysouw, M. & J. Good. 2007. Towards a comprehensive languoid catalogue. Presentation at the TOWARDS A COMPREHENSIVE LANGUOID CATALOGUE workshop at the MPI for Evolutionary Anthropology, Leipzig, 28 June 2007. Available at http://email.eva.mpg.de/~haspelmt/cat.html.

Feldpausch, B. 2005a. Bibliography of the SIL Papua New Guinea Branch 2001-2003 including materials produced by the Bible Translation Association of Papua New Guinea in cooperation with SIL Ukarumpa, Eastern Highlands Province, Papua New Guinea: Summer Institute of Linguistics .

Feldpausch, B. 2005b. Bibliography of the SIL Papua New Guinea Branch 2004 including materials produced by the Bible Translation Association of Papua New Guinea in cooperation with SIL Ukarumpa, Eastern Highlands Province, Papua New Guinea: Summer Institute of Linguistics .

Good, J. & C. Hendryx-Parker. 2006. Modeling contested categorization in linguistic databases. In *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*. Lansing, Michigan. http://www.linguistlist.org/emeld/workshop/2006/papers/GoodHendryxParker-Modelling.pdf.

Hammarström, H. 2008. Automatic annotation of bibliographical references with target language. In *Proceedings of MMIES-2: Wokshop on Multi-source, Multilingual Information Extraction and Summarization*, 57–64. ACL.

Lewis, P. M. (ed.). 2009. *Ethnologue: Languages of the world*. Dallas: SIL International, 16th edn. Online Version: http://www.ethnologue.com.

Linden, L. 2003. Bibliography of the Summer Institute of Linguistics Papua New Guinea Branch 1991-2000. Ukarumpa, Eastern Highlands Province, Papua New Guinea: Summer Institute of Linguistics.

AUTHOR CONTACT INFORMATION

Harald Hammarström
Centre for Language Studies
Radboud Universiteit
Postbus 9103
NL-6500 HD Nijmegen
The Netherlands

Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
Germany
h.hammarstrom@let.ru.nl

Sebastian Nordhoff
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
Germany
sebastian_nordhoff@eva.mpg.de

OSLa J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 45–62. (ISSN 1890-9639)
http://www.journals.uio.no/osla

THE NORDIC DIALECT CORPUS – A JOINT RESEARCH INFRASTRUCTURE

JANNE BONDI JOHANNESSEN

Department of Linguistics and Nordic Studies, University of Oslo

ABSTRACT

The paper describes the Nordic Dialect Corpus as of June 2010. The corpus (see Johannessen et al. 2009) is steadily growing, and new features are constantly added, so the version we describe is that of June 2010, while the corpus work has funding for another two years. The corpus is a tool that combines a number of useful features that together makes it a unique and very advanced resource for researchers of many fields of language studies. The corpus is web-based and features full audio-visual representation linked to transcriptions and translations.

[1] INTRODUCTION

In this paper, we describe the Nordic Dialect Corpus[1]. The corpus (see Johannessen et al. 2009) is steadily growing, and new features are constantly added, so the version we describe is that of June 2010, while the corpus work has funding for another two years. The corpus has a variety of features that combined makes it an advanced tool for language researchers. These features include: Linguistic contents (dialects from five closely related languages), annotation (tagging and two types of transcription), search interface (advanced possibilities for combining a large array of search criteria and results presentation in an intuitive and simple interface), many search variables (linguistics-based, informant-based, time-based), multimedia display (linking of sound and video to transcriptions), display of results in maps, display of informant details (number of words and other information on informants), advanced results handling (concordances, collocations, counts and statistics shown in a variety of graphical modes, plus further processing). Finally, and importantly, the corpus is freely available for research on the web. We give examples of both various kinds of searches, of displays of results and of results handling.

[2] WHY THE NORDIC DIALECT CORPUS WAS DEVELOPED

The Nordic Dialect Corpus was developed after a need for research material was voiced by members of the NORMS (Nordic Centre of Excellence in Micro-comparative Syntax) and the ScanDiaSyn (Scandinavian Dialect Syntax) networks.

The overarching goal for these researchers is to study the dialects of the North-Germanic languages, i.e., the Nordic languages spoken in the Nordic countries, as dialects of the same language. The languages are closely related to each other, and three of them are mutually intelligible (Norwegian, Swedish and Danish), as are two others (Faroese and Icelandic). All of them have some mutual intelligibility with each other if we consider written forms.

Studying the dialects only within the confines of each national language was therefore considered to be misguided from a theoretical and principled point of view. Second, doing research across dialects over such a big area, covering six countries (Denmark, Faroe Islands, Finland, Iceland, Norway, and Sweden), would be almost impossible if each researcher should get hold of relevant data on their own.

Third, the research in NORMS and ScanDiaSyn focusses on syntax – in which case data of many different kinds were necessary. Questionnaires for specific phenomena were needed (but will not be discussed in this paper), and recordings of spontaneous speech as it is used in ordinary conversations were very important. The latter need is satisfied by the Nordic Dialect Corpus.

[3] DESCRIPTION OF THE CORPUS

[3.1] *Linguistic contents and numbers*

The corpus contains dialect data from the national languages Danish, Faroese, Icelandic, Norwegian, and Swedish. It is steadily growing, since new recordings are still being done, or planned, while other recordings are in various stages of finishing. At the moment, it contains speech data from approximately 525 informants with 1.8 million words, unevenly spread between the five countries. Eventually, this will rise to around 600 informants. The numbers for the corpus as of today are given in Table 1.

Due to differences in the financing of the data collection in the different countries, the data are less uniform than one might have wanted ideally. (Some record-

[1] The Nordic Dialect Corpus is the result of close collaboration between the partners in the research networks Scandinavian Dialect Syntax and Nordic Centre of Excellence in Microcomparative Syntax. The researchers in the network have contributed in everything from decisions to actual work ranging from methodology to recordings, transcription, and annotation. Some of the corpus (in particular, recordings of informants) has been financed by the national research councils in the individual countries, while the technical development has been financed by the University of Oslo and the Norwegian Research Council, plus the Nordic research funds NOS-HS and NordForsk.

	Informants	Places	Words
Denmark	75	14	229 909
Faraoe Islands	19	5	48 427
Iceland	4	1	10 287
Norway	301	94	1 200 120
Sweden	126	40	299 86
Total	525	154	1 788 609

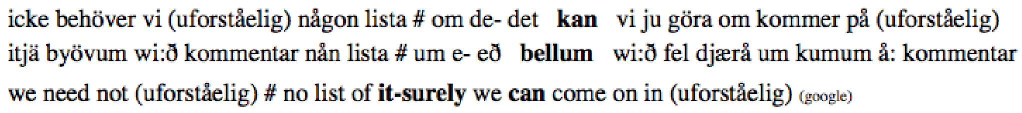
TABLE 1: Corpus contents by June 2010

ings and transcriptions were done for this corpus, while others were already done, such as most of the Swedish ones, which were generously given us by the earlier project Swedia 2000.)

Some recordings, such as those from Norway, the Swedish dialect of Övdalian and the Danish dialect of Western Jutlandic, have two kinds of recordings per informant: one semi-formal interview (informant and project assistant), and one informal conversation between two informants. Some dialects have recordings of both young and old informants, while others are only represented by old ones. Some dialects are represented by both old and new recordings, where old ones are generally around fifty years old. Some dialects have been recorded by audio only, while others have been recorded by both audio and video. All the dialects have recordings of informants belonging to both genders. Most importantly, however, all the recordings represent spontaneous speech.

[3.2] *Annotation: transcription and tagging*

All the dialect data have been transcribed by at least one transcription standard, and this work has been done for the most part in the individual countries: Each dialect has been transcribed by the standard official orthography of that country. (For Norwegian, which has two standard orthographies, Bokmål was chosen since there exist important computational tools for this variant.) In addition, all the Norwegian dialects and some Swedish ones have also been transcribed phonetically[2]. For the Norwegian dialects and the Övdalian Swedish ones that have two transcriptions, the first transcription to be done was in each case the phonetic one, and then the phonetic transcription was translated to an orthographic transcription via a semi-automatic dialect transliterator developed for the project. The fact that there are two transcriptions for dialects that are very different from the standard national orthography makes it possible to search with both transcriptions in the corpus, and present search results in both, as illustrated below for the Swedish dialect of Övdalian in Figure 1. This figure also shows the translation by Google, which is provided as a service in the corpus results presentation.



icke behöver vi (uforståelig) någon lista # om de- det **kan** vi ju göra om kommer på (uforståelig)
itjä byövum wi:ð kommentar nån lista # um e- eð **bellum** wi:ð fel djærå um kumum å: kommentar
we need not (uforståelig) # no list of **it-surely** we **can** come on in (uforståelig) (google)

FIGURE 1: Two transcriptions for Övdalian and a Google translation.

The Text Laboratory at the University of Oslo has the responsibility for the further technical devopment, including tagging. The whole corpus will be grammatically tagged with POS and selected morpho-syntactic features language by language. So far, the Norwegian data have been tagged, while the transcribed texts from the other languages are in the process of being tagged now. Tagging speech data is different from tagging written data. Speech contains disfluencies, interruptions and repetitions, and there are rarely clear clause boundaries (Allwood, Nivre and Ahlsén 1989, Johannessen and Jørgensen 2006). This is usually reflected in the transcription of speech, which generally does not contain clause boundaries or sentential markers such as full stops and exclamation marks (Jørgensen 2008, Rosén 2008). Any tagger developed for written language will therefore be difficult to use directly for spoken language. (Though Nivre and Grönqvist 2001 did this, on a material different from ours).

The Norwegian speech tagger was developed for the NoTa Corpus (Norwegian speech corpus – Oslo part). Søfteland and Nøklestad (2008) describe how the corpus was first tagged with the Oslo-Bergen tagger for written Norwegian (Hagen et al. 2000), and then trained with a TreeTagger (Schmid 1994) on the resulting, manually corrected file. The TreeTagger gained an accuracy of 96.9%. This tagger has then been used unchanged for the dialect corpus, under the assumption that the speech as represented in the dialects and in Oslo are sufficiently similar once they are all transcribed by the same transcription standard. The Swedish tagger has been trained in the same way. A written language TnT tagger developed by Sofie Johansson Kokkinakis (2003) has been applied to the Swedish dialect transcriptions (their standard orthographic version). After having been manually corrected and retrained, a spoken language Swedish statistical HunPos tagger has been developed at the Text Laboratory[3]. For Faroese, we have used a Faroese constraint grammar tagger developed for written language (Trosterud 2009), and manually corrected the results[4].

[2] The Norwegian phonetic transcription follows that of Papazian and Helleland (2005). The transcription of the Övdalian dialect follows the Övdalian orthography standardised in 2005 by the *Råðdjärum* (The Övdalian Language Council).

[3] The manual corrections of the Swedish tagger were done by Piotr Garbacz, and the tagger was developed by André Lynum, both at the Text Laboratory, UiO.

[4] The manual corrections of the Faroese tagger were done by Remco Knooihuizen for the Text Laboratory, UiO.

[3.3] *Search Interface*

The corpus uses an advanced search interface and results handling system, Glossa (Nygaard 2007, Johannessen et al. 2008). The system allows for a large variety of search combinations making it possible to do very advanced and complex searches, even though the interface is very simple, with pull-down menus, and boxes that expand only when prompted by the user. The corpus search system Corpus Work Bench (Christ 1994, Evert 2005) is used, so that the simple corpus queries are translated to regular expressions before querying – something that is invisible to the user.

Several of the features in the search interface and the results display follow suggestions by participants in ScanDiaSyn and NORMS.

Searching for lemmas and part of words: For those parts of the corpus that are tagged and lemmatised, it is possible to search for the lemma only. This way we get all inflected forms of one lexeme. This feature is very useful when there is suppletion in the stem of the word. For example, search for the Norwegian lemma *gås* ('goose') will give the results *gås, gåsa, gjess, gjessene* (various combinations of number and definiteness).

The same box where the user can write a full search word or a lemma can also be used to write part of a search word. This way the user can, for example, search for a particular suffix. In Figure 2, the user has searched for the suffix –*ig*, which can be found in Norwegian, Swedish, and Danish.

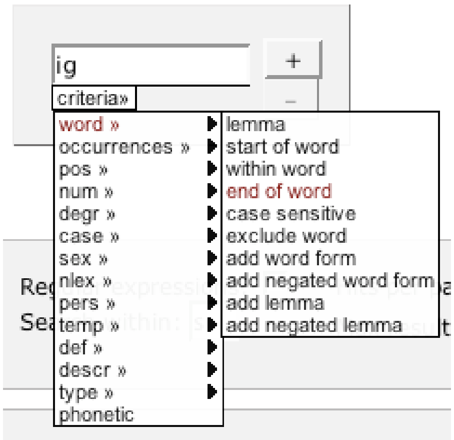


FIGURE 2: Search for suffix -*ig*

Notice that since nothing else was specified, this search would query the whole corpus, all the languages. In Table 2 we can see some of the many hits for the frequent adjectival suffixes –*ig* and -*lig* in the mainland Nordic languages, and a couple of occurrences of words containing the same sequence of letters in the insular Nordic languages (not representing these suffixes, however).

Freq.	Word	Translation	Language
7	særlig	especially	No, Da
7	farlig	dangerous	No, Sw, Da
7	þannig	thus	Ice
7	kjedelig	boring	No
6	väldig	very	Sw
5	rigtig	right	Da
5	otrolig	unbelievable	Sw
4	konstig	strange	Sw
1	sjómannaslig	sailor-like	Fa

TABLE 2: Some results from the –*ig* search

Searching for more than one word: In order to specify a search for more than one word, the user clicks on the plus sign in the first box, which gives one more box, with the possibility of specifying a number of words in between (Figure 3).

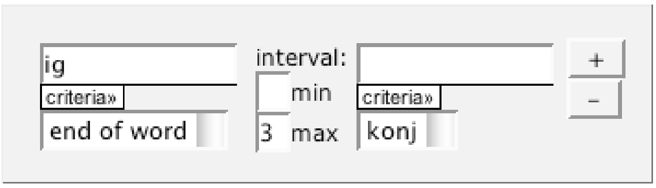
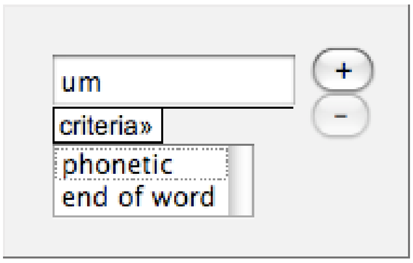


FIGURE 3: Searching for two words

The illustration shows a search for a word ending in –*ig* separated by at most three words from a conjunction to the right.

Searching for part of speech: The tagged part of the corpus can also be queried directly by part-of-speech tags. This is exemplified in Figure 3, where the second word is specified to be a conjunction. The user can choose whether a search word is specified by a word form (or part of one) and a part of speech or both. The pull-down menus in Figure 2 exemplify many of the search options that are available for a word.

Phonetic querying: The user can choose to query the corpus by giving a phonetically specified string. This works only for the dialects that have two transcriptions (cf. section 4.2). An example of a situation in which this is useful will be where we want to query person-number inflection on verbs. Here, tagging will not help, since each tagger is trained on the standard orthographic version of the texts, and person-number inflection is only a dialect feature. Searching for this feature in Övdalian, we can simply write for example the 1pl suffix as it is (Figure 4):


um

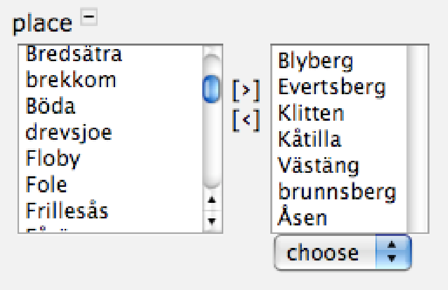
criteria»

phonetic
end of word

FIGURE 4: Searching in phonetic mode

This will give results that would have been impossible to get from the orthographic transcriptions. We refer to Figure 1, where the dialectal *bellum* ('can' 1pl) is represented by the standard *kan* ('can').

Informant-based querying: There are a number of ways to query the corpus in addition to the linguistics-based ones that we have seen above. All the details that are known about each informant are also searchable in the search interface. Thus, it is possible to specify as search criteria: age, sex, recording year, place of residence, country, region and area. In Figure 5, we show how we can choose individual places from the complete list, to be able to query only the informants from these places, which happen to be the area of Älvdalen in Sweden.


place

Bredsätra
brekkom
Böda
drevsjoe
Floby
Fole
Frillesås

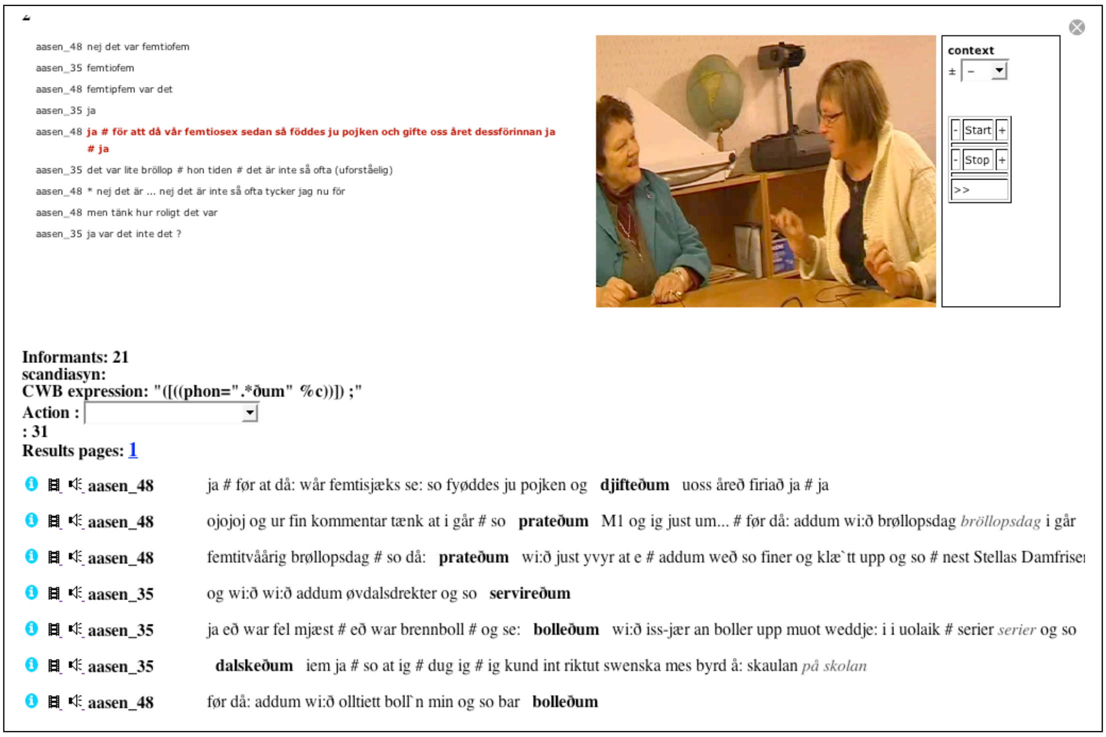
Blyberg
Evertsberg
Klitten
Kåtilla
Västäng
brunnsberg
Åsen

FIGURE 5: Delimiting the corpus by choosing some places from the full list

[3.4] *Display of search results*

Each search in the corpus gives a standardised view of the results in the form of a classical KWIC concordance. The results can be viewed in a number of additional ways which we will present below.

Multimedia display: The corpus includes transcribed speech from five countries and spans four decades. Some of the speech was recorded using a tape recorder and later mp3 recorder, and some was recorded by videocamera. The search result is accompanied by a clickable symbol to show the audio and video of that particular speech sequence. This is illustrated in Figure 6 below.

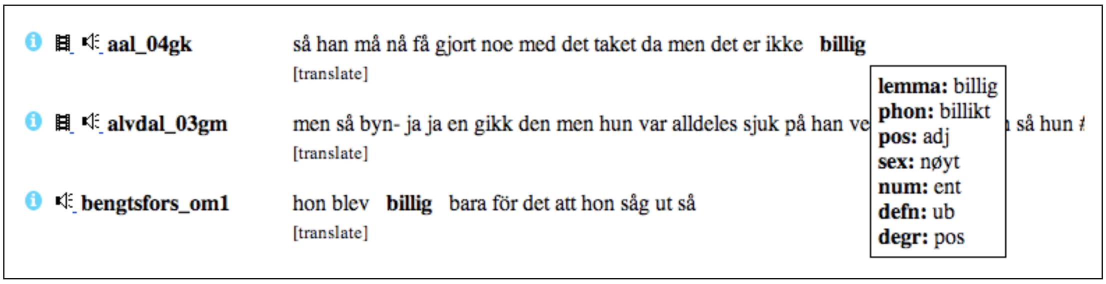

aasen_48 nej det var femtiofem
aasen_35 femtiofem
aasen_48 femtipfem var det
aasen_35 ja
aasen_48 **ja # för att då vår femtiosex sedan så föddes ju pojken och gifte oss året dessförinnan ja # ja**
aasen_35 det var lite bröllop # hon tiden # det är inte så ofta (uforståelig)
aasen_48 * nej det är ... nej det är inte så ofta tycker jag nu för
aasen_48 men tänk hur roligt det var
aasen_35 ja var det inte det ?

Informants: 21
scandiasyn:
CWB expression: "([((phon=".*ðum" %c))]) ;"
Action :
: 31
Results pages: 1

aasen_48 ja # før at då: wår femtisjæks se: so fyøddes ju pojken og **djifteðum** uoss åreð firiað ja # ja
aasen_48 ojojoj og ur fin kommentar tænk at i går # so **prateðum** M1 og ig just um... # før då: addum wi:ð brøllopsdag *bröllopsdag* i går
aasen_48 femtitvåårig brøllopsdag # so då: **prateðum** wi:ð just yvyr at e # addum weð so finer og klæ`tt upp og so # nest Stellas Damfriser
aasen_35 og wi:ð wi:ð addum øvdalsdrekter og so **servireðum**
aasen_35 ja eð war fel mjæst # eð war brennboll # og se: **bolleðum** wi:ð iss-jær an boller upp muot weddje: i i uolaik # serier *serier* og so
aasen_35 **dalskeðum** iem ja # so at ig # dug ig # ig kund int riktut swenska mes byrd å: skaulan *på skolan*
aasen_48 før då: addum wi:ð olltiett boll`n min og so bar **bolleðum**

FIGURE 6: The multimedia results window

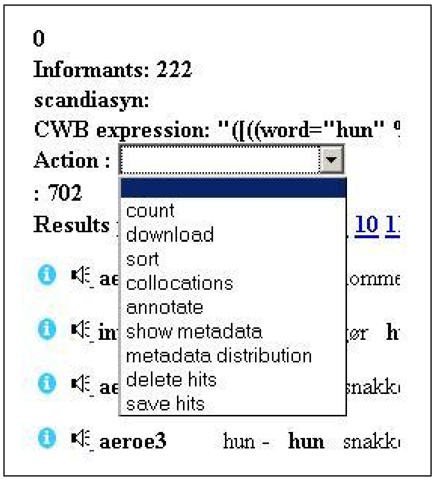
Display of transcriptions and tagging: For those linguistic variants that have two transcriptions, either transcription can be chosen for displaying the result. The grammatical tags and the phonetic transcription of each standard orthographic word are visible in a box when mousing over the text (Figure 7).


aal_04gk så han må nå få gjort noe med det taket da men det er ikke **billig**
[translate]
alvdal_03gm men så byn- ja ja en gikk den men hun var alldeles sjuk på han ve så hun #
[translate]
bengtsfors_om1 hon blev **billig** bara för det att hon såg ut så
[translate]

lemma: billig
phon: billikt
pos: adj
sex: nøyt
num: ent
defn: ub
degr: pos

FIGURE 7: A window shows all information for each word that is moused over

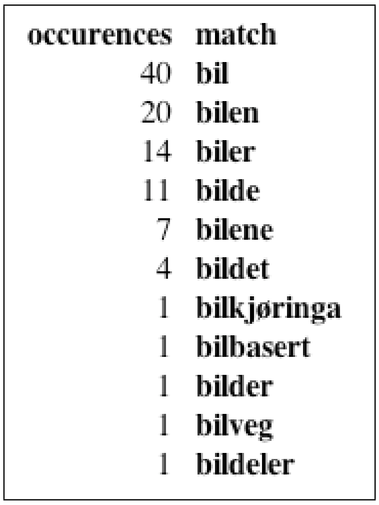
Action menu: On the results page there is an Action menu with a selection of choices for further displaying of results and results handling (the latter of which will be presented in section [3.6]). The functionalities that follow in this subsection are choices in this menu (Figure 8).



0
Informants: 222
scandiasyn:
CWB expression: "([((word="hun"
Action :
: 702
Results

FIGURE 8: Action menu in results window

Count: Choosing the Count option gives the search results as a list of all the hits sorted by frequency. In Figure 9, a bit of a list is shown as a result of the search for nouns starting with *bil-* in Norwegian.



occurences	match
40	bil
20	bilen
14	biler
11	bilde
7	bilene
4	bildet
1	bilkjøringa
1	bilbasert
1	bilder
1	bilveg
1	bildeler

FIGURE 9: Some nouns beginning with *bil-* (‘car’)

The count results can be shown in a number of ways, such as histograms and pie charts.

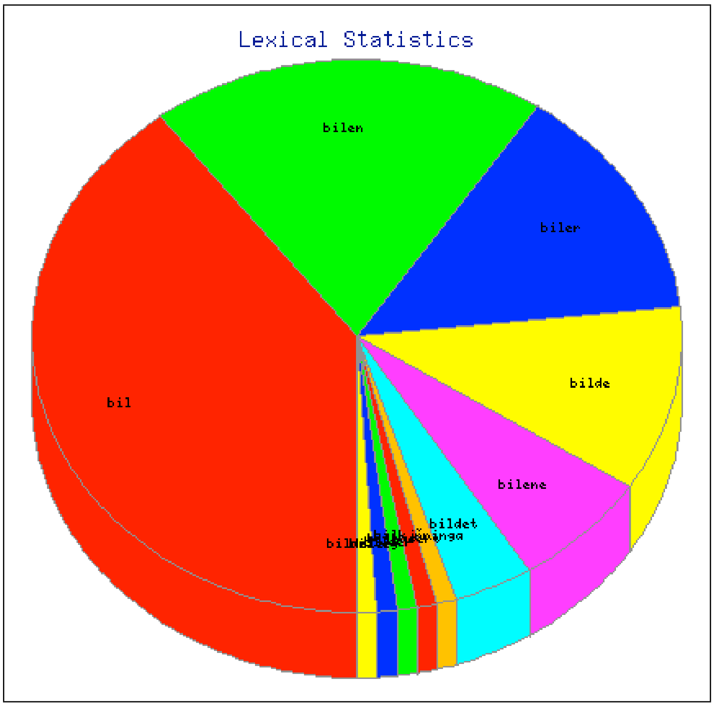
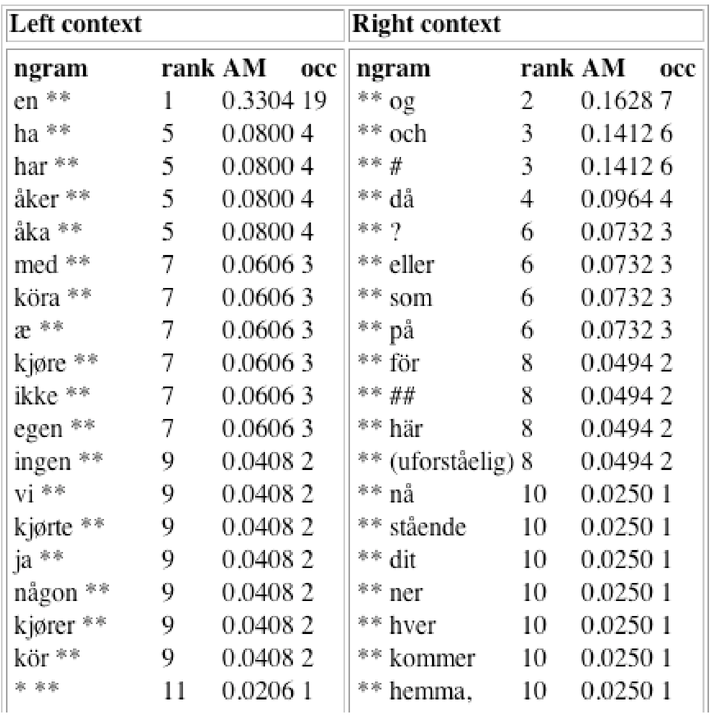


FIGURE 10: The same information as in Figure 9.

Sort: The results are by default sorted according to the geographical residence of the informants. However, they can be displayed in many other ways as well. The most useful ones are perhaps those that sort the matches by the next word to the right or left.

Collocations: The results can be shown as collocations according to many different statistical measurements such as dice coeffiency, log-likelihood ratio etc., with a choice between neighbouring bigrams and trigrams. The example in Figure 11 illustrates the collocations for the word *bil* 'car', used in the three mainland Nordic countries. The value of this choice is clearly illustrated in the example in Figure 11; the frequencies of the collocations are the same independently of language.

Maps: Recently an option of displaying the search results on maps (using Google Maps technology) has been added. Since one search can cover a variety of results, for example when one orthographic word covers many different phonetic varieties, an additional option has been added in which each variety can be selected independently. In the map in Figure 12 the different phonetic varieties of the negation are displayed in the right-hand column, giving the user the choice to choose one or more and have them independently shown on the map. The orthographic variety has been displayed by a neutral dot covering all pronunciations.



Left context				Right context			
ngram	rank	AM	occ	ngram	rank	AM	occ
en **	1	0.3304	19	** og	2	0.1628	7
ha **	5	0.0800	4	** och	3	0.1412	6
har **	5	0.0800	4	** #	3	0.1412	6
åker **	5	0.0800	4	** då	4	0.0964	4
åka **	5	0.0800	4	** ?	6	0.0732	3
med **	7	0.0606	3	** eller	6	0.0732	3
köra **	7	0.0606	3	** som	6	0.0732	3
æ **	7	0.0606	3	** på	6	0.0732	3
kjøre **	7	0.0606	3	** för	8	0.0494	2
ikke **	7	0.0606	3	** ##	8	0.0494	2
egen **	7	0.0606	3	** här	8	0.0494	2
ingen **	9	0.0408	2	** (uforståelig)	8	0.0494	2
vi **	9	0.0408	2	** nå	10	0.0250	1
kjørte **	9	0.0408	2	** stående	10	0.0250	1
ja **	9	0.0408	2	** dit	10	0.0250	1
någon **	9	0.0408	2	** ner	10	0.0250	1
kjører **	9	0.0408	2	** hver	10	0.0250	1
kör **	9	0.0408	2	** kommer	10	0.0250	1
* **	11	0.0206	1	** hemma,	10	0.0250	1

FIGURE 11: Some collocations for *bil* ‘car’.

[3.5] *Displaying information on informants*

There are two ways of finding information on the informants. **Via results page:** Each concordance line has an information symbol on its very left. Clicking on this symbol reveals information on the informant in question: informant code, sex, age group, country, place, number of words, recording year, and recently we have also included a map for his/her home place, see Figure 13.

Via search page: There is a button called “Show Texts”, which shows information on which informants are included in a particular query. For example, if the user wants to query the corpus on Swedish data only, (s)he can press this button and immediately see how many informants are represented in the selection, how many words each informant has uttered etc., and this information can also be sorted by category to present for example number of words in a descending order. This way, we can see how different the informants are in this respect. For example, one old man from Skreia, Norway, utters 1,300 words during his session, while another old man, from nearby Stange, utters more than 6,400 words.

[3.6] *Further processing of results*

Deleting or choosing some results: In a corpus search it is often the case that the user gets more results than intended. Sometimes the search expression just was

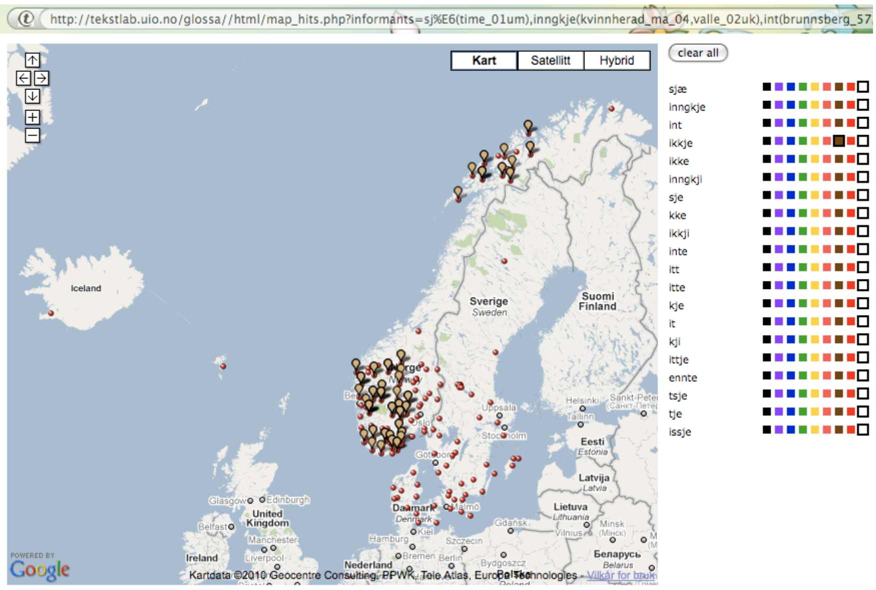


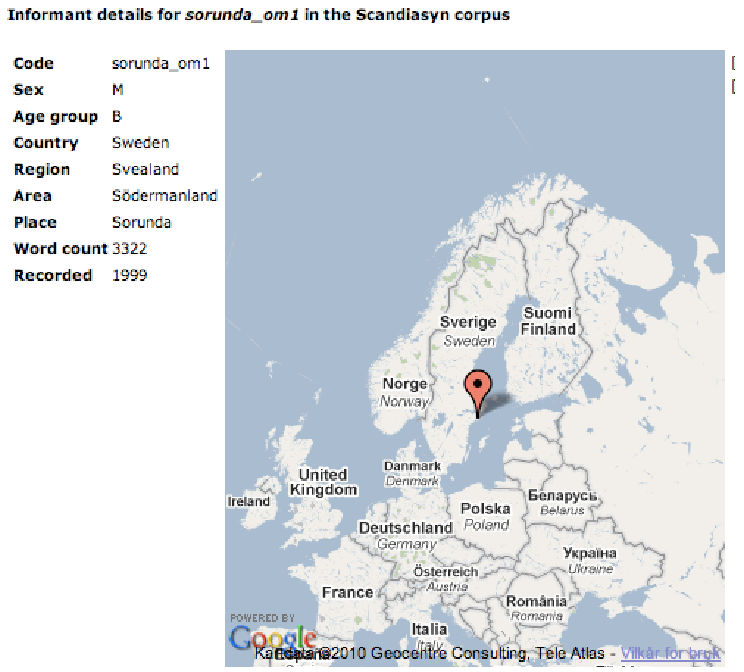
FIGURE 12: A map showing all the places that have hits (all the dots) for the orthographic forms of the negation 'not'. The column on the right can be specified for a phonetic variant. Here the phonetic form *ikkje* has been chosen. It should be noted that parts of North Norway have not yet been included in the corpus.

not good enough, which can best be corrected by a new and more precise search. However, sometimes it is impossible to formulate better search criteria, whether it is because there is too much homonymy in the corpus, or because it just is not annotated for all imaginable research features. Let us use a simple example: We want to find all and only the occurrences of the 3sgF pronoun ('she') used as a determiner, followed by any word, and then a noun. This search will give a lot of unwanted hits that we want to remove. We can then choose the Delete option from the Action menu and get Figure 14.

Notice in the figure that by having chosen the Delete option, the results come with a little box on the left hand side. In this box we tick the examples that we want to remove. If we suspected that there would only be a few examples that were appropriate for our research, we could instead have used the Choose option, which functions in the same way, but where ticking a box would mean to keep that result and delete the unticked ones.

Annotating results: The individual researcher often needs to further annotate the results, for example according to pronunciation of certain sounds or words, or specific syntactic patterns. In Figure 15, we have chosen to annotate the examples by two categories: Demonstrative or Other.

The annotations can be edited and saved as annotation sets, for later reuse



Informant details for *sorunda_om1* in the Scandiasyn corpus

Code	sorunda_om1
Sex	M
Age group	B
Country	Sweden
Region	Svealand
Area	Södermanland
Place	Sorunda
Word count	3322
Recorded	1999

FIGURE 13: Information that appears in the search results window

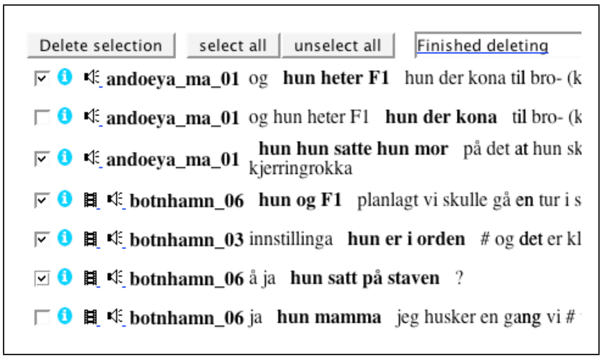
with other results.

Saving and downloading results: All results can be saved and/or downloaded, whether we choose the raw results or those that we have further processed by deletion, choice or annotation. By saving we get the opportunity to look at the results later, and with exactly the same possibilities for further processing and displaying of results in the corpus interface. Downloaded results, on the other hand, are not thus available in the corpus system, but can be imported as for instance tab-separated text.

[4] COMPARISON WITH OTHER DIALECT CORPORA

There are some other dialect resources on the web, but there are to our knowledge few or no available web-based dialect multimedia corpora for other languages. One interesting resource is *Sounds familiar? Accents and Dialects of the UK*. It contains information on British dialects, and recordings of the dialects with transcripts, all presented via a web map. However, it is pedagogical, and not aimed at researchers. For example, there is no search option in the transcripts and no grammatical annotation.

The Scottish Corpus of Text and Speech contains 4 million words, 20% of which is spoken texts, provided with orthographic transcription, synchronised with the audio or video. It is not grammatically annotated and is not representative. How-



☑ **andoeya_ma_01** og **hun heter F1** hun der kona til bro- (k

□ **andoeya_ma_01** og hun heter F1 **hun der kona** til bro- (k

☑ **andoeya_ma_01** **hun hun satte hun mor** på det at hun sk kjerringrokka

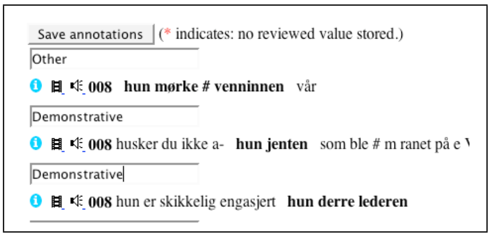
☑ **botnhamn_06** **hun og F1** planlagt vi skulle gå **en** tur i s

☑ **botnhamn_03** innstillinga **hun er i orden** # og **det** er kl

☑ **botnhamn_06** å ja **hun satt på staven** ?

□ **botnhamn_06** ja **hun mamma** jeg husker en gang vi #

FIGURE 14: Results window with Delete option



(* indicates: no reviewed value stored.)

Other

008 **hun mørke # venninnen** vår

Demonstrative

008 husker du ikke a- **hun jenten** som ble # m ranet på e

Demonstrative

008 hun er skikkelig engasjert **hun derre lederen**

FIGURE 15: Results window with Annotate option

ever, it has a nice search interface.

The British National Corpus contains 10 million words of spoken English, which have been categorised into 28 different dialects. However, it says in their own search interface distribution that this categorisation is unreliable. Further, as a dialect corpus, the BNC has limited value, since it is not represented with audio, and the speech is transcribed orthographically.

The DynaSand web-based dialect database consists of information on various syntactic features and their distribution geographically in the Netherlands and Belgium. It contains recorded material from the project's questionnaire sessions, but the conversations contain to a large extent read sentences and meta-linguistic discussions, and less spontaneous speech.

The Spoken Dutch Corpus is transcribed orthographically, some of it also phonetically, and it is morphologically tagged. It contains spoken standard Dutch, not dialect data, and is not available by a web-interface.

The Corpus of French Phonology (La phonologie du français contemporain: usages, variétés et structure – PFC) is a web-based corpus of spoken French from across the Francophone world. It is searchable both phonologically and w.r.t. informant characteristics, and has transcriptions linked to sound.

There might be web-based dialect corpora for other languages, but information about these is hard to find, and they do not seem to be available on the web.

One such corpus under development is Corpus of Estonian Dialects. Another is Spoken Japanese Dialect Corpus (GSR-JD), available on DVD. Finally we should mention a small dialect corpus of Norwegian (Talesøk). It contains audio and transcriptions, and is available on the web.

There are some general web-based speech corpora that do not focus on dialect classification. For an overview of some Northern European ones, and their state of art w.r.t. topics like technical solutions and audio-visual availability, we refer to Johannessen et al. (2007).

Finally, we would like to mention that Paul Thompson at the University of Reading had a posting at Corpora List on November 30 2008 asking for information on corpus projects in which the developers have linked digital audio and/or video files to the transcripts, to allow access to the precise segment(s) of the audiovisual files that relates to a part of the transcript. In his summary of 15 responses there was only one dialect corpus – our own Nordic Dialect Corpus.

[5] CONCLUSION

We have presented the first version of the Nordic Dialect Corpus. It contains nearly 1.8 million words of Nordic dialects as spontaneous, not manuscripted, conversations. Most of them have been collected recently, but we have also included some old speech data. The Nordic Dialect Corpus has an advanced interface for searching and results handling. It is already a great resource for dialect researchers and linguists interested in the Nordic languages. The next version of the corpus will contain more dialect data. Part-of-speech taggers adapted for speech will be developed for alle the languages, and all present and future texts will be tagged.

[6] ACKNOWLEDGEMENTS

In addition to participants in the ScanDiaSyn and NORMS networks, I would like to thank three anonymous NODALIDA-09 reviewers for valuable comments on an earlier version on this paper. I would also like to thank various funding bodies for funding the technical part of this project: the University of Oslo (both the Humanities Faculty and the central Research Department), The Norwegian Research Council, NordForsk and NOS-HS. In addition the national research councils in Norway, Sweden, Denmark and Iceland have contributed to the projects NorDiaSyn, SweDiaSyn, DanDiaSyn and IceDiaSyn, as has the project Swedia 2000, which has contributed a lot of Swedish recordings, and Norsk Ordbok 2014, which has contributed many months worth of recordings for the old Målførearkiv material that is included in the corpus.

Finally, I would like to emphasise that developing a corpus of the kind that I have described here is team work. I want to mention the cooperation of Øystein

Alexander Vangsnes (University of Tromsø), Tor Anders Åfarli (the Norwegian University of Science and Technology, Trondheim) and the staff at the Text Laboratory, especially Kristin Hagen, Signe Laake, Anders Nøklestad and Joel Priestley.

REFERENCES

Allwood, Jens, Joakim Nivre & Elisabeth Ahlsén. 1989. Speech management - On the nonwritten life of speech. In *Gothenburg Papers in Theoretical Linguistics*. University of Gothenburg.

Christ, Oliver. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. In *COMPLEX'94*. Budapest.

Evert, Stefan. 2005. *The CQP Query Language Tutorial*. Institute for Natural Language Processing, University of Stuttgart. URL www.ims.unistutgart.de/projekte/CorpusWorkbench/CQPTutorial.

Hagen, Kristin, Janne Bondi Johannessen & Anders Nøklestad. 2000. A Constraint-based Tagger for Norwegian. In Carl-Erik Lindberg & Steffen Nordahl Lund (eds.), *17th Scandinavian Conference of Linguistics*, Odense Working Papers in Lanugage and Communication 19, 31–48. University of Southern Denmark, Odense.

Johannessen, Janne Bondi & Kristin Hagen. 2008. *Språk i Oslo. Ny forskning omkring talespråk.* Novus Forlag, Oslo.

Johannessen, Janne Bondi, Kristin Hagen, Joel Priestley & Lars Nygaard. 2007. An Advanced Speech Corpus for Norwegian. In *NODALIDA Proceedings*, 29–36. Tartu: University of Tartu.

Johannessen, Janne Bondi & Fredrik Jørgensen. 2006. Annotating and Parsing Spoken Language. In Peter Juel Henriksen & Peter Rossen Skadhauge (eds.), *Treebanking for Discource and Speech*, 83–103. København: Samfundslitteratur.

Johannessen, Janne Bondi, Lars Nygaard, Joel Priestley & Anders Nøklestad. 2008. Glossa: a Multilingual, Multimodal, Configurable User Interface. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Paris: European Language Resources Association (ELRA).

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli & Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In Kristiina Jokinen & Eckhard Bick (eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4.*

Johansson, Sofie Kokkinakis. 2003. *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS*. Ph.D. thesis, Gothenburg University.

Jørgensen, Fredrik. 2008. Automatisk gjennkjenning av ytringsgrenser i talespråk. In Janne Bondi Johannessen og Kristin Hagen (ed.), *Språk i Oslo*. Novus Forlag.

Nivre, Joakim & Leif Grönqvist. 2001. Tagging a Corpus of Spoken Swedish. *International Journal of Corpus Linguistics* 6(1). 47–48.

Nygaard, Lars. 2007. *The glossa manual*. The Text Laboratory. URL www.hf.uio.no/tekstlab/glossa.html.

Papazian, Eric & Botolv Helleland. 2005. *Norsk talemål*. Høyskoleforlaget, Kristiansand.

Rosén, Victoria. 2008. Mot en trebank for talespråk. In Janne Bondi Johannessen & Kristin Hagen (eds.), *Språk i Oslo*. Novus Forlag.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Søfteland, Åshild & Anders Nøklestad. 2008. Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. In Janne Bondi Johannessen og Kristin Hagen (ed.), *Språk i Oslo*, 226–234. Novus Forlag.

Thompsom, Paul. 2008. Summary on Info of audio-visual corpora. *Corpora List* .

Trosterud, Trond. 2009. A constraint grammar for faroese. In Eckhard Bick, Kristin Hagen, Kaili Müürisep & Trond Trosterud (eds.), *Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing. NEALT Proceedings Series*, vol. 8, 1–7.

CORPORA AND WEB RESOURCES

Barbiers, S. et al (2006). Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND). Amsterdam, Meertens Institute.
http://www.meertens.knaw.nl/sand/

British National Corpus:
http://www.natcorp.ox.ac.uk/

Corpus Gesprochen Nederlands.
http://lands.let.kun.nl/cgn/ehome.htm

Nordic Dialect Corpus:
http://omilia.uio.no/glossa/html/index_dev.php?corpus=scandiasyn

NoTa Corpus (Norwegian speech corpus – Oslo part)
http://www.tekstlab.uio.no/nota/oslo/

La phonologie du français contemporain : usages, variétés et structure (PFC)
http://www.projet-pfc.net/pfc-recherche

Sounds familiar?
http://www.bl.uk/learning/langlit/sounds/index.html

Scottish Corpus of Text and Speech.
http://www.scottishcorpus.ac.uk/

Spoken Japanese Dialect Corpus (GSR-JD)
http://research.nii.ac.jp/src/eng/list/detail.html#GSR-JD

Swedia 2000.
http://swedia.ling.gu.se/

Talesøk.
http://helmer.aksis.uib.no/talekorpus/Hovedside.htm

Text Laboratory, UiO:
http://www.hf.uio.no/tekstlab/English/index.html

AUTHOR CONTACT INFORMATION

Janne Bondi Johannessen
Department of Linguistics and Nordic Studies
University of Oslo
P.O. Box 1102 Blindern
N-0317 Oslo
Norway
jannebj@iln.uio.no

OSLa J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 63–74. (ISSN 1890-9639)
http://www.journals.uio.no/osla

THE EDISYN SEARCH ENGINE

JAN PIETER KUNST AND FRANCA WESSELING
Amsterdam

ABSTRACT

Edisyn (European Dialect Syntax) is a project on dialect syntax funded by the European Science Foundation. It runs at the Meertens Institute in Amsterdam from September 2005 until September 2010 (partially extended till March 2012). It aims at achieving two goals. One is to establish a European network of (dialect) syntacticians that use similar standards with respect to methodology of data collection, data storage and annotation, data retrieval and cartography. The second goal is to use this network to compile an extensive list of so-called doubling phenomena from European languages/dialects and to study them as a coherent object. One of the deliverables of the Edisyn project is a web-based search engine to search different linguistic corpora simultaneously and show the combined search results. This search engine is able to make differently structured databases comparable. Although the initial set up of the Edisyn project was to create similar standards for dialectal databases, in practice this has proven to be an unfeasable goal since most databases have a different structure and enrichment (we will come back to this below). Consequently, the Edisyn search engine has been created according to a more pragmatic philosophy and is able to handle databases of various structures.

[1] INTRODUCTION

The Edisyn project focuses on doubling phenomena in various languages. Since these phenomena primarily occur in non-standard varieties, their existence has gone largely unnoticed in the linguistic literature. Recent literature on Dutch dialects (SAND project) has revealed a wealth of doubling phenomena that do not appear in Standard Dutch. See for instance in the cases in (1)-(4) below[1].

(1) Subject pronoun doubling and subject agreement doubling:
Ze peiz-n da-n ze ziender rijker zij-n
They think-3PL that-3PL they they richer are-3PL
’They think that they are richer.’

[1] PART = participle, PL = plural, 2 = second person, 3 = third person

(2) WH-word doubling:
Wel denkst wel ik in de stad ontmoet heb
Who think-2PL who I in the city met have
’Who do you think I met in the city?’

(3) Participial morphology doubling:
Zol hee dat edane hemmn ekund
Would he that done-PART have could-PART
’Could he have done that?’

(4) Auxiliary doubling:
K-em da gezegd gehad
I-have that said-PART had-PART
’I have said that.’

Through the investigation of non-standard varieties, doubling phenomena can be adequately researched. The project therefore greatly enhances the empirical basis of syntactic research. Cross-linguistic comparison of doubling phenomena will enable us to test or formulate new hypotheses about natural language and language variation. By investigating doubling phenomena we are able to detect the pervasiveness and limitations hereof. The Edisyn project seeks to answer the question whether there are any limitations as to what kind of linguistic categories can be subjected to doubling. Furthermore an explanation is sought for any such restrictions. These answers will not only contribute to the characterization of micro-variation but will in turn have implications on how we look both at meso-variation (e.g. OV word order versus VO order) and macro-variation (e.g. polysynthetic versus non-polysynthetic).

To enhance cross-linguistic research on non-standard varieties a search engine - the so called Edisyn search engine- has been created enabling comparative research on dialect data of different languages. Until recently most dialectological work focused on variation within the non-standard varieties of one language, the availability of the Edisyn search engine, however, enables the investigation of dialects of various languages. At the moment of writing(March 2010) five databases containing data on non-standard varieties of a specific language have been combined within a single interface. The unified search interface allows the user to search different European linguistic corpora of dialect transcriptions simultaneously and shows the combined search results on a single results page. Searching for text strings and textual patterns should be possible, though this kind of search is of limited value when searching text across different languages. At the moment a basic search for strings is possible. The problems that arise when attempting to connect linguistic corpora are outlined below.

[2] LINGUISTIC DESIGN OF THE EDISYN SEARCH ENGINE

[2.1] *Introduction*

Every database has its own, specific structure. This is due to various reasons. First of all, a dialect database differs according to the (type of a) language. The content of a database is dependent upon syntactic and morphological properties of a language. If a language has case marking, for example, the values hereof will be part of the tag set of a database. If a language does not assign case, these features will be absent in the database.

Second, the structure of a database depends on the kind of data that has been gathered. If the data consists of elicited speech different choices are made with respect to the structure of the database than if the data concerns, say, spontaneous speech. In a database containing elicited speech, the dialect data can be lined up with the question (or test) sentences, whereas this is not possible with spontaneous speech. In the latter case the data will be more difficult to parse and specific decisions need to be made concerning the desired way of presenting the data.

Furthermore, the theoretical views of the linguist(s) can alter the outlook of a database. If one is working within a generative framework the tags that are assigned may be theory-dependent. In the ASIt database (Italian dialects)[2] for example, the tag *raising* is assigned to certain parts of speech types. This tag is used to indicate verbs that do not assign an external theta role such as *appear* and *seem*, whereby the semantic subject of the lower clause verb is syntactically realised as a constituent of the higher clause. This term is highly theory-dependent for it is not used in non-generative frameworks.

Also, the set up of a database is influenced by the subject matter of a research. If data is collected within a research project focused on the order of verbs in subordinate clauses, for instance, the ordering of verbs (and possibly other part of speech types) will be tagged. Other syntactic or morphological phenomena may then receive less attention/marking.

A crucial factor in the make up of a database is the kind of enrichment a database contains. A database may only have raw recordings, or these recordings may be lined up per sentence so that small parts of a conversation can be listened to. Furthermore, these recordings may be tagged with part of speech tags -per word- or keywords may be assigned to an entire phrase, which are in turn database specific. A database may also have both enrichment at the word level and contain syntactic parsing. In addition, the data may be translated into English, this can be done word by word, or apply to entire phrases.

[2] The ASIt database is available at http://asis-cnr.unipd.it/

The databases also differ with respect to the quality of the enrichment, that is, the assignment of tags can be very detailed or less thorough. In addition, the subject matter that is tagged can vary, for instance, question sentences can be enriched with linguistic tags (in the case of elicited data), or the answers can be tagged, or both question sentences and answers may be tagged. Also, databases will be dissimilar in the extensiveness of the English translations. Finally, the metadata is an important aspect that differs per database. This kind of information is often present to a limited extent, or not at all. Ideally, every database would provide information specifying the period in which the data has been gathered, the location(s) where the research has been undertaken, the kind of data that is presented in the database, the age of the informants, the people in charge of the research and database and their affiliation, et cetera. However, often none of these details are further specified, let alone in a similar fashion.

In summary, databases vary from one another in the following respects:

(1) Type of data: type of language, elicited data versus spontaneous speech.

(2) Enrichment of data: part of speech tags / syntactic labels / linguistic keywords / English glosses / a combination hereof.

(3) Quality of the enrichment: the data is meticulously tagged / the data is tagged in a more general manner.

(4) Quantity of the enrichment: only answers are tagged / only question sentences are tagged / both questions and answers are tagged / neither is tagged.

(5) Metadata: information specifying the circumstances in which the data has been gathered is absent or the databases provide this information to a different degree.

In the attempt of making different databases interoperable via one search engine, these differences need to be considered. Ideally, each database would have similar standards with respect to the factors mentioned above, this is however never the case. Nevertheless, it is feasible to create a search engine that queries various databases which contain data that has been tagged and glossed, despite their external differences. This has been done in the Edisyn project, resulting in the Edisyn search engine. Via this search engine it is possible to search on the basis of part of speech tags and on the basis of strings of words, the latter search option being of course highly language specific.

Note that in the development of the search engine, the Edisyn team has no desire to change the configurations of any of the component databases. The aim of the search engine is simply to provide a tool via which it is possible to search dialect data of various languages through a single interface. Each database retains its own tag set and can be consulted individually at all times.

Abbreviation:	Category:
V	Verb
N	Noun
D	Determiner
Pron	Pronoun
A	Adjective
Adv	Adverb
Conj	Conjunction
Negmrk	Negation marker
P	Adposition
C	Complementizer
Part	Particle
Intj	Interjection

TABLE 1: Part of speech categories used in Edisyn search engine

[2.2] *The Edisyn Tag Set*

The first and perhaps most important step in connecting different databases is to equalize the different tag sets. Within the Edisyn project we have constructed a general tag set containing part of speech categories and linguistic features (this division will be elaborated upon below), as shown in Table 1 and 2. This tag set can be 'translated' to many different tag sets (note that the Edisyn tag set is dynamic and can be adjusted according to the needs of a database developer).

A *category* refers to commonly used parts of speech such as Verb, Noun, Adjective, etc. These can be combined with *features* such as 'singular' which results in a specific tag, such as a singular noun. A category can be combined with any and as much feature(s) as desired. Thus any tag can be created. However, not every query will generate a result because not every database has assigned the same tags to their data. This is clearly communicated to the user of the Edisyn search engine. That is, if a query has no result, the user is informed that the tag in question has not been assigned in the individual database.

Categories cannot be combined with other categories. Thus, the category Noun can be combined with the feature 'nominative case', but not with the category Adjective. It is possible to search for a sequence of categories, for instance Noun followed by Adjective. Tags can be either adjacent to each other -the default setting- or with an optional gap (zero or more words in between the tags).

The home page of the search engine consists of an overview of the databases that can be consulted. By clicking on the box next to each database, the database selected will be included in the following query. It is also possible to search each database in its original layout, the link next to each database connects the user

Abbreviation:	Feature:	Abbreviation:	Feature:
ab	abessive case	m	masculine
abl	ablative case	mesacl	mesaclisis
acc	accusative case	mod	modality
act	active	neg	negative
ad	adessive case	neut	neuter
add	additive case	nom	nominative case
all	allative case	num	numeral
art	article	partit	partitive case
asp	aspect	pass	passive
aux	auxiliary	past	past tense
caus	causative	perf	perfective
cl	clitic	pers	personal
com	comitative case	pl	plural
comp	comparative	poss	possessive (case)
coord	coordinating	post	postposition
dat	dative case	pp	past articiple
def	definite	prep	preposition
dem	demonstrative	pres	present tense
dim	diminutive	presp	present participle
el	elative case	procl	proclisis
encl	enclisis	quant	quantitative
erg	ergative	recipr	reciprocal
es	essive case	refl	reflexive
f	feminine	rel	relative
fin	finite	sg	singular
foc	focus (marker)	subord	subordinating
fut	future tense	sup	superlative
gen	genitive case	term	terminative case
ger	gerund	tr	translative case
ill	illative case	trans	transitive
imp	imperative	unacc	unaccusative
in	inessive case	unerg	unergative
indef	indefinite	1	first person
infin	infinitive	2	second person
inst	instrumental case	3	third person
inter	interrogative (=wh)		
intrans	intransitive		

TABLE 2: Linguistic features used in Edisyn search engine

directly to that database. When using the Edisyn search engine the tag set described above and in Tables 1 and 2 is to be used, if one is querying an individual database the tag set of that specific database is of course employed.

After one or more database(s) has/have been selected one can start creating a tag, this is done by adding one or more features to a category, as described above. It is also possible to search for a category or feature by itself. When the appropriate tag has been selected, the search engine will present the results available for the selected database(s).

Note that when a query has been performed with the Edisyn search engine, the results contain the tags of the individual database. For example, if one wants to know if dialects of Portuguese and dialects of Dutch both have a way of marking a verb in the present tense for second person singular, one adds these databases to the search by selecting them. Then, one drags the category Verb to the search field, followed by the features ’pres’, ’2’ and ’sg’. By clicking on *search* the query is started and the results will be shown. These results contain -in this example- the dialect sentences in Portuguese, with the tags provided by the Cordial-Sin database, and the data in Dutch with the tags used in the SAND database. These tags are easily interpreted by the user for all the tags used in the various databases are explained in a glossary.

The results are based on the conversion of the tag set of the Edisyn search engine to the tag set of each database. That is, at the backend, the tag used in the search engine is connected to the corresponding tag in each database. Every category and every feature has a corresponding tag in each of the databases, for instance, in the example above the Edisyn tag ’V(fin,pres,2,sg)’ is linked to the Portuguese tag ’V-P-2S’.

With the Edisyn tag set available many databases can be interconnected via the search engine for each tag set can be translated into so called Edisyn tags. Again, we want to stress that we do not make any changes to the individual databases; we leave the structure and tag set of each database completely intact. Via the conversion of the Edisyn tag set to the tag sets of the databases it is possible to search various (dialect) databases at the same time, enabling a cross-linguistic comparison of dialect data.

[2.3] *Note on English Glosses*

It is of importance to add English glosses to a database, for this will enhance the accessibility of the search engine and it will allow more researchers to use the database. Most researchers will have (some) knowledge of the language and its dialects (s)he is working on, but this need not be the case for the other dialect databases which have been made interoperable in the Edisyn search engine. With the addition of English glosses however, all the dialect data is made comprehensible for every (English speaking) linguist, and may trigger their interest. By making the content accessible to everyone in the field more research on dialects may even be stimulated.

Currently the database on Dutch dialects (SAND) contains English glosses, that is, there is a translation available for every word that is used in this database. The Cordial-Sin corpus (on Portuguese dialects) is working on the implementation of a word by word translation into English. Within the Nordic Dialect Corpus there is a possibility of translating every sentence by Google Translate. The other databases do not have an application to display the dialect data in English. This is work to be done in the future.

[3] IMPLEMENTATION OF THE SEARCH ENGINE

[3.1] *Ideal Architecture for a Search Engine*

The ideal architecture of a search engine would, in our view, be a distributed one: each research group hosting, maintaining, and being responsible for its own corpus, and exposing its search interface via a web service, i.e. an interface for computer programs, as opposed to human users, to access the corpus. The central search engine then calls the different corpora via these web service interfaces, and shows the combined results on its own results page. In practice, such an ideal architecture is difficult to realize. Some linguistic corpora do not have a search interface as such, but are simply made available as downloadable text files. Other corpora do have a web-based search interface, but strictly one for human users. In those cases the research groups responsible for the corpora usually do not have the resources to add the needed features to their existing corpora.

In those cases we opted for the pragmatic solution of hosting copies of the corpora locally on our own server. Of course, this makes problems like handling updated versions of corpora more complicated than in a web service-based solution, but that is a necessary trade-off in this situation, because otherwise there would not be a search engine at all. In the case of the Nordic Dialect Corpus we access the corpus remotely, at the moment of writing not yet via a true web service but by doing normal http requests with a *curl* library and 'screen scraping' the returned pages with results. We hope to convert this system to a real web service connection in the future.

But, even if, in many instances, we have to work with locally hosted corpora out of necessity, we still built the search engine using a web service architecture, with *localhost* URLs for the corpora. This makes it relatively easy to switch to a remote web service for a corpus if the opportunity arises: change the URL to point to the remote host instead of localhost. It is unlikely that the interface will be exactly the same as the one we created ourselves for our localhost web services, so probably some additional fine-tuning will be needed, but that will certainly be less work than converting a platform-specific local connection for a corpus to a web service connection.

[3.2] *Current State of the Search Engine*

At the moment an experimental version of the Edisyn Search Engine is online at `http://www.meertens.knaw.nl/edisyn/searchengine/`, with five corpora included: SAND (Syntactic Atlas of the Dutch Dialects, Dutch, Meertens Institute), CORDIAL-SIN (Syntax-oriented Corpus of Portuguese Dialects, Portuguese, University of Lissabon), ASIt (Syntactic Atlas of Northern Italy, Italian, University of Padua), EMK (Estonian Dialect Corpus, Estonian, University of Tartu) and NDC (the Nordic Dialect Corpus, Scandinavian Languages, ScanDiaSyn). With the exception of the Nordic Dialect Corpus, all corpora are hosted locally at the Meertens Institute.

Searching for POS tags is enabled via a central Edisyn tag set (visible in the 'tags' menu on the search page; see fig. 1 for a screenshot). The user can search for complete tags, partial tags, or features. For each corpus, there is an XML file which translates the tags from the central tag set into the native tag set of the corpus. So the central search engine is quite 'shallow' and does not know anything about the tag sets of the corpora it uses, in turn, the participating corpora only see search requests with their native tag sets and do not know anything about the Edisyn tag set. This set up makes it possible to add new corpora to the search engine without affecting the existing system.

[3.3] *Technical Details of the Search Engine*

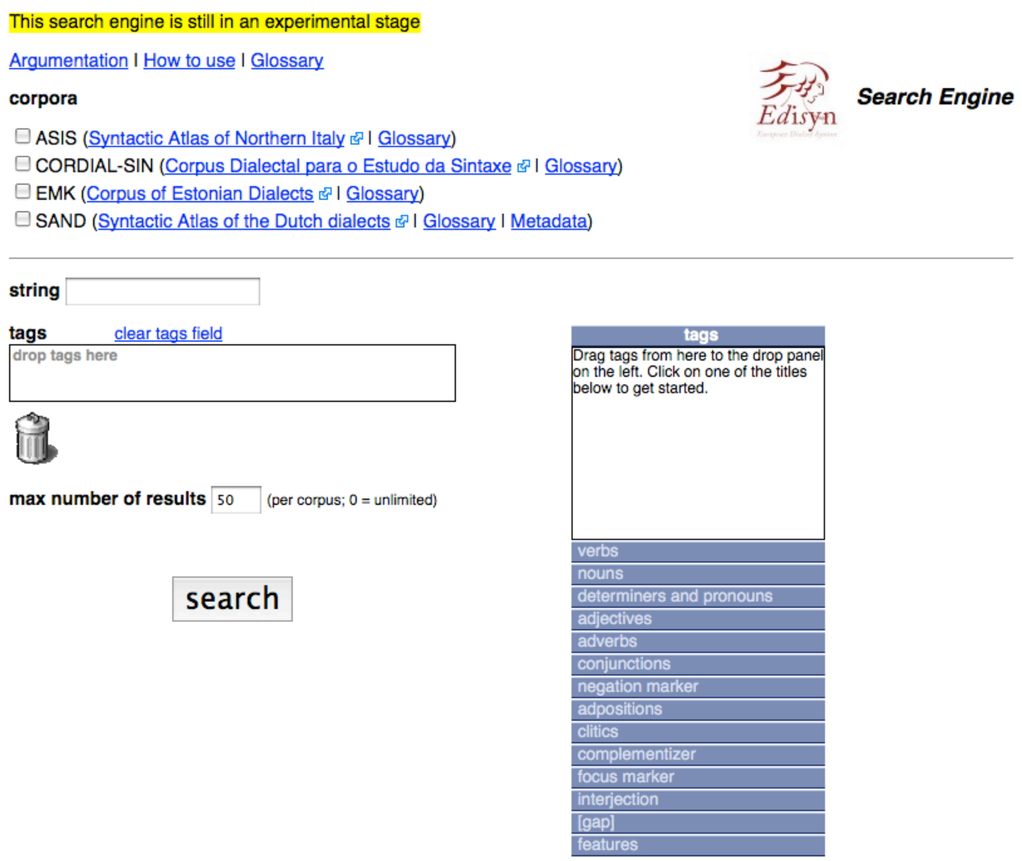
User Interface

As mentioned before, the Edisyn search engine is web based and should work in any reasonably modern browser. The user interface consists of standard XHTML pages enriched with JavaScript via the JQuery library. We use JQuery to create a drag-and-drop interface for constructing search queries, in order to make the potentially tedious process of entering POS tags in queries as streamlined as possible; and we use an AJAX interface (also provided by JQuery) to the server to avoid unnecessary page reloads.

Server-side Technologies

The Edisyn Search Engine is written in object-oriented PHP. The web page containing the search form is created by a class EdisynPage. This class creates the search form and checks if it has been submitted; if it is, it fetches the search results and adds them to the page; if not, it just shows the form.

Fetching the results is done by instantiating search classes for each checked corpus, called `Edisyn_Search_<corpusname>`. As their name implies, these classes are corpus-specific; they are child classes of an abstract class `Edisyn_Search` which contains general, non-corpus-specific methods and properties. The knowledge about how the searches are performed is encapsulated in the search classes; the EdisynPage class just feeds the form data to the search classes and calls a



This search engine is still in an experimental stage

Argumentation | How to use | Glossary

corpora

□ ASIS (Syntactic Atlas of Northern Italy | Glossary)
□ CORDIAL-SIN (Corpus Dialectal para o Estudo da Sintaxe | Glossary)
□ EMK (Corpus of Estonian Dialects | Glossary)
□ SAND (Syntactic Atlas of the Dutch dialects | Glossary | Metadata)

Drag tags from here to the drop panel on the left. Click on one of the titles below to get started.

max number of results 50 (per corpus; 0 = unlimited)

FIGURE 1: Screen shot of Edisyn search engine

getResults() method on them.

[4] FUTURE PROSPECTS FOR THE SEARCH ENGINE

The Edisyn Search Engine in its current state is not finished. We list some features and enhancements which will be added in the future in this section.

[4.1] *Mapping*

An option to show search results on a map will be added in the future. The groundwork is already there: almost all of the data which is hosted locally at the Meertens Institute is enriched with geographical coordinates, as is the Nordic Dialect Corpus, so enhancing the search results to include geographic locations is not a difficult problem. This will provide the user with the possibility to show the data from different corpora combined on a single map of Europe. We plan to use Google Maps as the web mapping solution to display these data.

[4.2] *Additional Corpora*

Some corpora which we plan on adding in the near future are: the Afrikaans Variation Project (Mark de Vos, Rhodes University), Slovene Dialectical Syntax (Marko Hladnik, University of Utrecht), Diversion in Dutch DP Design (DiDDD, University of Utrecht) and Freiburg English Dialect Corpus (FRED, University of Freiburg). In the distant future we also hope to give acces to Lauseopin Arkisto on Finnish dialects (Kotus (Research Institute for the Languages of Finland)), COSER on Spanish dialects (Corpus Oral y Sonoro del Español Rural, Autonomous University of Madrid) and a database of Breton dialects (ARBRES, Melanie Jouitteau). At the moment data on Basque dialects are being gathered at the University of Bayonne (IKER), which will also be made interoperable by the Edisyn search engine.

It is our aim to add as many databases as possible, the requirements for a suitable database being rather limited, namely having reliable and useful data on any (European) dialect, which has been tagged and preferably contains English glosses.

[4.3] *Clarin*

The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable. Standards for data and metadata for language resources are being developed in the CLARIN project. We plan to adhere to these CLARIN standards to preclude the Edisyn project being an isolated effort. For further information about CLARIN, see http://www.clarin.eu.

One of the standards which are being developed within CLARIN is the so called ISOcat category set. This allows linguists to tag their data with a dataset which has been approved by the ISOstandard (ISO 12620 provides a framework for defining data catagories according to the ISO/IEC 11179 family of standards). At this moment we are modifying the Edisyn tag set according to the standard of the ISOcat categories. This will lead to a more unified way of tagging which will make dialect databases more comparable.

Finally, we will develop and implement more user-friendly applications along the way. That is, more differentiated search options will be added and other enhancements which prove to be useful, will be put into effect.

REFERENCES

Barbiers, S. & H. Bennis. 2007. The syntactic atlas of the dutch dialects. a discussion of the choices in the SAND-project. In K. Bentzen & Ø. Vangsnes (eds.), *Nordlyd*, vol. 34, 53–72.

Barbiers, S., L. Cornips & J.P. Kunst. 2007. The syntactic atlas of the dutch dialects (SAND): A corpus of elicited speech and text as an on-line dynamic atlas. In

J. Beal, K. Corrigan & H. Moisl (eds.), *Creating and digitizing language corpora: Vol. 1, synchronic database*, 54–90. Hampshire: Palgrave-Macmillian.

Barbiers, S. et al. 2006. Dynamic syntactic atlas of the dutch dialects (Dynasand). URL http://www.meertens.knaw.nl/sand/. Amsterdam, Meertens Institute.

Benincà, P. & C. Poletto. 2007. The asis enterprise: a view on the construction of a syntactic atlas for the northern italian dialects. In K. Bentzen & Ø. A. Vangsnes (eds.), *Nordlyd*, 34, 35–52.

Johannessen, J. B., J. Priestley, K. Hagen, T.A. Åfarli & Ø. A. Vangsnes. 2009. The nordic dialect corpus - an advanced research tool. In K. Jokinen & E. Bick (eds.), *NEALT proceedings series*, vol. 4. Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA.

St.Laurent, S., J. Johnston & E. Dumbill. 2001. *Programming web services with xml-rpc*. Sebastopol: O'Reilly.

AUTHOR CONTACT INFORMATION

Jan Pieter Kunst
Meertens Institute
P.O. Box 94264
NL-1090 GG Amsterdam
The Netherlands
janpieter.kunst@meertens.knaw.nl

Franca Wesseling
Meertens Institute
P.O. Box 94264
NL-1090 GG Amsterdam
The Netherlands
franca.wesseling@meertens.knaw.nl

OSLa J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 75–95. (ISSN 1890-9639)
http://www.journals.uio.no/osla

AGGREGATE ANALYSIS OF VOWEL PRONUNCIATION IN SWEDISH DIALECTS

THERESE LEINONEN
Society of Swedish Literature in Finland

ABSTRACT

In this paper an aggregate analysis of vowel pronunciation in Swedish dialects is proposed by means of multidimensional scaling (MDS). The Gap statistic showed that no statistically significant partitioning of Swedish dialects can be made based on vowel pronunciation, which means that the dialects form a true linguistic continuum. Vowels recorded by 1,170 speakers at 98 sites were analyzed acoustically with principal components of Bark-filtered spectra, and the linguistic distances between varieties were computed as the Euclidean distance of the acoustic variables. The MDS analyses showed that the dialect areas that can be detected based on vowel pronunciation in modern rural varieties of Swedish largely correspond to the traditional Swedish dialect division and divisions of regional varieties of Standard Swedish. The results also show a large-scale ongoing dialect leveling. The change is largest in many central parts of the language area close to the biggest cities, while the dialects in more peripheral areas are relatively stable.

[1] INTRODUCTION

The Swedish dialects have gone through massive leveling in the latter half of the 20th century (Hallberg 2005; Thelander 2005). Due to a societal change from rural societies to urban life style, including migrations from the countryside to towns and cities, traditional rural dialects have disappeared in many parts of the language area and been replaced by (regional varieties of) Standard Swedish. In this leveling process especially morphological, syntactical and lexical variation has decreased profoundly. Phonetic and prosodic features are assumed to have been preserved to a larger degree (Engstrand et al. 1997).

Figure 1 on page 77 shows the traditional division of Swedish rural dialects by Wessén (1969). According to Wessén (1969, 12–13), the rural Swedish dialects have formed a continuum without any sharp dialect borders. In this continuum, however, Wessén identified six main dialect areas: South Swedish dialects (*sydsvenska mål*), Götaland dialects (*götamål*), Svealand dialects (*sveamål*), Norrland dialects (*norrländska mål*), Gotland dialects (*gotländska mål*), and Finland-Swedish dialects (*östsvenska mål*).

Elert (1994) proposed a division of the regional varieties of Standard Swedish. He based the division mainly on sentence intonation and differences in vowel pronunciation. The geographic division by Elert largely resembles the classification of the traditional rural dialects by Wessén (1969).

In order to make instrumental analysis of variation in modern rural varieties of Swedish possible, dialect data was recorded at a large number of sites around year 2000 in the project SweDia (see below, Section [2]). Based on SweDia data, Bruce (2004) classified Swedish dialects according to intonational variation. The intonational parameters of the model were focal accentuation, phrasing, word accentuation and compounding. Seven distinct dialect regions were identified, largely corresponding to the ones found by Wessén and Elert.

Both rural Swedish dialects and regional varieties of Standard Swedish vary a lot when it comes to vowel pronunciation, and vowels have been important for characterizing varieties of Swedish and classifying dialects (Bruce 2010, 102–103). Still, only very few instrumental analyses of vowels covering the whole Swedish language area exist. In this paper the variation in vowel pronunciation in modern rural varieties of Swedish is analyzed by means of acoustic analysis of vowels recorded at nearly 100 sites in Sweden and the Swedish-speaking parts of Finland. The speakers represent two different age-groups which makes the study of language change in apparent time possible.

Aggregating methods introduced by the dialectometric research tradition (Nerbonne 2009) are used for identifying geographic dialect areas and studying ongoing change in this paper. Aggregation originally developed as a more objective alternative to the isogloss method for defining dialect areas and exploring dialect continua. While isoglosses are chosen subjectively by the researcher, aggregating methods can incorporate much larger amounts of data simultaneously and also deal with conflicting signals in the data.

[2] DATA

The data analyzed in this paper were recorded within the SweDia project (Eriksson 2004), a collaboration between the universities of Lund, Stockholm and Umeå. The data were gathered during the period 1998–2001 at 98 rural sites (see Figure 1 on the next page) in the Swedish language area. The locations were chosen to represent the dialectal situation in the Swedish language area by being balanced geographically and with respect to population density. At each location recordings were made with around twelve speakers: three older women, three older men, three younger women and three younger men. The older speakers were in the approximate age range of 55–75 years when the recordings were made, and the younger speakers were 20–35 years. The total number of speakers included in the analyses in this paper is 1,170.

Vowel segments were elicited with mono- or bi-syllabic words. To keep the

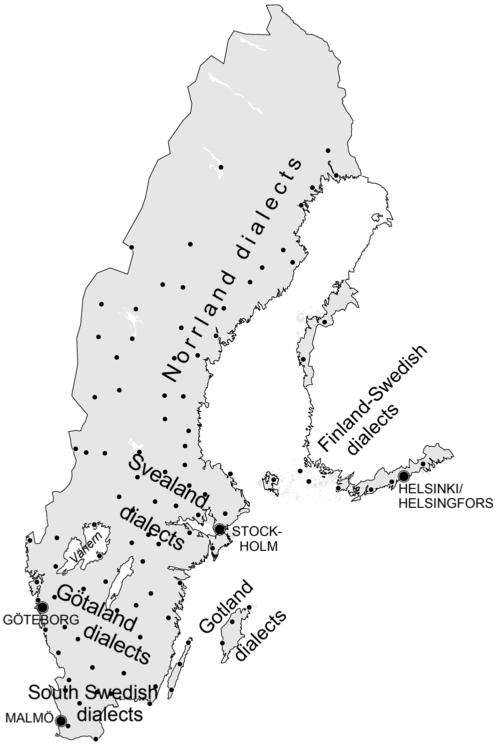


FIGURE 1: The traditional Swedish dialect areas according to Wessén (1969), and the 98 data sites of the current study. The four biggest cities in the area are included as reference points in the map.

phonetic context of the vowels as stable as possible, the target vowels were surrounded by coronal consonants. For the current study the vowels from the stressed syllables in the following 19 words were used (Standard Swedish vowel pronunciation in square brackets): *dis* [iː], *disk* [ɪ], *dör* [œː], *dörr* [œ], *flytta* [ʏ], *lass* [a], *lat* [ɑː], *leta* [eː], *lett* [e̞], *lott* [ɔ], *lus* [ʉː], *lås/låt* [oː], *lär* [æː], *lös* [øː], *nät* [ɛː], *sot* [uː], *särk* [æ], *söt* [øː], *typ* [yː].

The data set includes all of the Standard Swedish long vowel phonemes. In addition the allophonic variants of /ɛː/ and /øː/ ([æː] and [œː] which occur only before /r/), and a few vowels which reflect language historical developments are included. For example, the words *lös* and *söt* have the same vowel phoneme in Standard Swedish, but historically the former originates from a diphthong /au/, which has been monophthongized in most varieties of Swedish, while the latter one is an original monophthong. Some of the Swedish dialects have preserved these vowels as two different phonemes. The phoneme /oː/ was elicited with the word *lås* in some sites and the word *låt* in others. Of the Standard Swedish short vowel phonemes a few are missing because they had not been consistently elicited at all sites for the database. Even though a few short vowel phonemes are missing, the data should be able to give a good picture of how the Swedish dialects relate to each other with respect to vowel pronunciation. The variation in vowel pronunciation across varieties of Swedish has been described to be more prominent in long vowels than in short vowels (Elert 2000, 38).

A few sites which are well known for their well-preserved divergent rural dialects (Orsa and Älvdalen) have been excluded from the current analysis. This is because they were considered so different from other dialects during the SweDia fieldwork that a completely different word list was used for eliciting vowel sounds in these dialects. Hence, there is no directly comparable vowel data from these dialects in the SweDia database.

The recordings were made in the speakers' homes or other familiar places in order to make the participants feel comfortable and to make the use of the local vernacular feel natural. A lapel microphone and a portable DAT-recorder were used, and the recordings were done at 48 kHz sample rate and 16-bit amplitude resolution. Before analysis the data were downsampled to 16 kHz/16 bit. Each speaker repeated the words 3–5 times. The recordings were annotated and the vowels manually segmented within the SweDia project.

[3] METHODS

[3.1] *Acoustic analysis*

All vowel segments were filtered with Bark filters up to 18 Bark at nine equidistant sampling points within each segment with a window length of 13 ms. The first sampling point was at 25% of the total vowel duration and the last at 75%. The Bark-filtered spectra were level-normalized for every 13 ms sample so that the levels add up to 80 dB. A representation in Bark filters gives a good perceptual representation of vowels, because the Bark scale corresponds to the critical bandwidth of human hearing.

After the Bark-filtering each sampling point of each vowel pronunciation was characterized by the levels (in dB) in the consecutive frequency filters. Since the speakers had repeated the 19 different vowels 3–5 times the average levels of these repetitions were computed for each speaker.

The Bark-filtered data were subsequently reduced to two principal components (PCs) by principal component analysis (PCA) with varimax rotation. PCA is a data reduction technique that aims at reducing a larger number of variables into a smaller set of components by combining variables that correlate with each other (Tabachnik & Fidell 2007). Reducing a filter bank representation of speech samples to PCs leads to an articulatory meaningful configuration. PC1 is related to vowel height, while PC2 is related to tongue advancement. PCs of Bark-filtered vowel spectra have been shown to correlate highly with formant measurements by Jacobi (2009) and Leinonen (2010). The method is well-suited for large-scale analysis of vowel pronunciation because it is more reliably automatable than formant measurements, which always need to be manually corrected.

Following Jacobi (2009) only point vowels were used for building up the PCA, which means that all articulatory dimensions are represented equally. For the Swedish data the vowels transcribed as [iː], [æː], [ɑː]/[aː] and [uː] in the database were chosen for calculating the loadings in the initial phase of the PCA.

Because of the anatomical/physiological differences between male and female vocal tracts (men have on average longer vocal tracts than women) the formants related to vowel production have higher frequencies in female voices than in male voices (Peterson & Barney 1952). These differences can be normalized for by applying PCA separately to data from male speakers and females speakers (Leinonen 2010). Vowels produced by male speakers were analyzed in the frequency range 2–17 Bark, while vowels produced by female speakers were analyzed at 3–18 Bark.

In the initial phase of the PCA point vowels from 230 women and 230 men were used for computing loadings. Subsequently PC scores were computed for all vowels of all speakers. In the male analysis PC1 explains 41.1% of the total variance in the data and in the female analysis PC1 explains 41.4% of the variance. PC2 explains 36.3% (male) respectively 36.8% (female) of the variance. Together the two extracted components explain nearly 80% of the variance in both analyses. For a more detailed description of the acoustic analysis see Leinonen (2010).

[3.2] *Analysis of dialectal variation*

Common aggregating techniques used in dialectometric research for exploring dialect areas and dialect continua are cluster analysis and multidimensional scaling (MDS). In both methods a distance matrix with the aggregate pairwise distances between all objects is used as input. MDS is a method for reducing complex distance data to interpretable low-dimensional representations, while cluster analysis produces partitions of the data. MDS is suitable for visualizing dialect continua, while cluster analysis detects dialect groups.

The Gap statistic can be used for estimating the number of significant clusters produced by any clustering algorithm (Tibshirani et al. 2001). Lundberg (2005) used the Gap statistic to estimate the number of significant clusters when grouping the Swedish dialects based on acoustic analysis of the vowel in the word *lat* (/ɑː/ in Standard Swedish) and found three significant clusters. The Gap statistic was applied to the present data set, and the analysis showed that there are no well separated clusters. This result indicates that the Swedish dialects form a true continuum when it comes to an aggregate analysis of vowel pronunciation.

Clustering methods could be applied to the data, but they are likely to produce unstable results, since any sharp division into subsets is not in agreement with the structure of the data. Therefore MDS was chosen for analyzing the dialectal variation in vowel pronunciation in this paper. The MDS plots (for example, Figure 2 on page 82) confirm that the Swedish language area can be best described as a dialect continuum when it comes to vowel pronunciation.

In MDS, original distances between objects are approximated in a low-dimensional space by an iterative algorithm (Jain & Dubes 1988). A number of algorithms for MDS have been proposed. In a study of the linguistic distances between varieties of Dutch, Heeringa (2004) measured the fitness of three different MDS procedures by correlating the original distances with the Euclidean MDS coordinate-based distances, and found that Kruskal's non-metric MDS gave the best results. In this paper Kruskal's non-metric MDS, as implemented in the RuG/L04[1] software, was used.

For the MDS the distances between the varieties were calculated as the average distance of the 19 vowels in the data set. First, the distance for each vowel between two varieties was calculated as the Euclidean distance of the acoustic variables of vowel quality (Equation 1), that is, the two PCs measured at nine sampling points within each vowel segment, starting at 25% of the total vowel duration and ending at 75%. Subsequently, the average of the vowel distances was calculated. For some groups of speakers a few of the vowels are missing, and the distances involving these objects are the average distances of a fewer number of vowels. At least 15 of the 19 vowels were recorded in each speaker group.

The equation below shows the Euclidean distance, where i ranges over the nine sampling points per vowel and x and y are either two different sites (Section [4.1]) or two different speaker groups (older or younger speakers at any of the sites, Section [4.2]):

[1] RuG/L04 – software for dialectometrics and cartography. By P. Kleiweg, University of Groningen. <http://www.let.rug.nl/kleiweg/L04/>

$$distance(x, y) = \sqrt{\sum_{i=1}^{9}((PC1_{xi} - PC1_{yi})^2 + (PC2_{xi} - PC2_{yi})^2)} \qquad (1)$$

When applying MDS to dialect data three dimensions generally explain at least around 90% of the total variance in the data (for example, Heeringa 2004; Prokić & Nerbonne 2008), which means that additional dimensions are not important for describing the dialectal variation. In MDS to three dimensions positions in a three-dimensional space are assigned to all varieties included in the analysis. One way of displaying the results is to plot the objects in a Cartesian coordinate system. The closer to each other two objects are in the coordinate system the smaller the linguistic difference. However, the interpretation of the results of MDS is facilitated if the results can be viewed in relation to geography. By using the `maprgb` function in the RuG/L04 software, which uses the RGB color model, the results of MDS can be displayed on maps (Nerbonne et al. 1999). By using the RGB color model all positions in the three-dimensional MDS space are translated to a distinct color. The amount of red represents the first dimension of the MDS, the amount of green the second dimension and the amount of blue the third dimension. Coloring the area surrounding each site on a map with the color corresponding to the position assigned by MDS links the results of MDS to geography.

[4] RESULTS

Below the results of MDS applied to two different divisions of the data are presented. In Section [4.1] the geographic variation is described by averaging over all speakers per site. In Section [4.2] the data is split into older and younger speakers per site in order to analyze the degree of language change in apparent time. Group means of the speakers were calculated for PC1 and PC2 respectively for each of the 19 vowels at the nine sampling points. Subsequently the Euclidean distances between varieties were calculated based on the group means, and the resulting distance matrix was analyzed with MDS.

[4.1] *Geographic dialect continuum*

In the first MDS analysis, average values for each site (that is, the averages of the approximately twelve speakers) were calculated for the acoustic variables before measuring the linguistic distances between sites and applying MDS. The distance matrix comprised the pairwise distances between all 98 sites.

Figure 2 on the next page shows the results of the MDS in a two-dimensional coordinate system where the color of the dot represents the third dimension. One dimension explains 81.4% of the variance, two dimensions 93.6% and three dimensions 96.3%. It is clear that the first dimension already explains a very large part

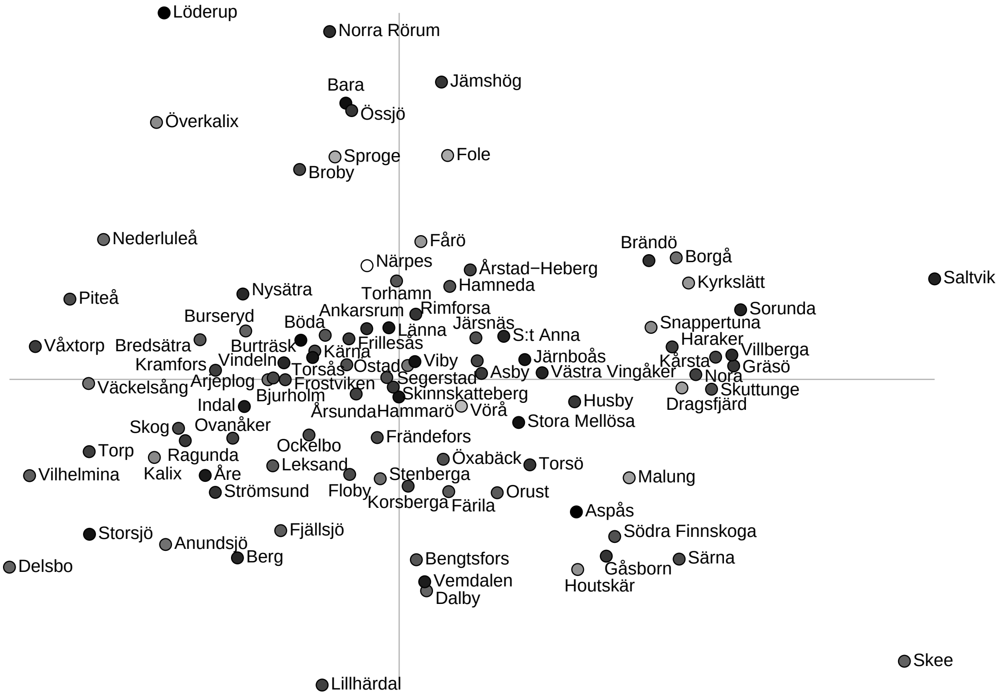


FIGURE 2: Results of MDS to 3 dimensions of the linguistic distances between sites. The 1st dimension is represented by the x-axis, the 2nd dimension by the y-axis and the 3rd dimension by the gray-scale color of the dot. Because of the density of sites close to the origin, a few labels have been omitted.

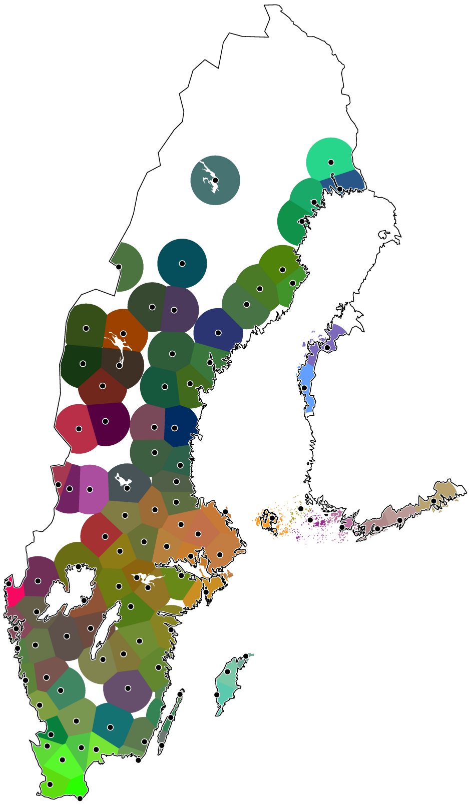


FIGURE 3: Results of MDS to 3 dimensions of the linguistic distances between sites, displayed with the RGB color model.

of the variance. Using more than three dimensions would only mean a small improvement of the variance explained. The plot does not show any clear clusters of sites, but there is only one big cloud, which supports the result of the Gap statistic that the dialects form a true continuum. The plot shows a concentration of sites close to the origin and a more sporadic distribution in the peripheries.

In the first dimension, sites in Svealand (mainly the province Uppland) and on the Finnish south coast have high values, while mainly sites in Norrland have low values. Sites with high values on the second dimension are the South Swedish ones, but also the ones on Gotland. The third quadrant (negative values on the two first dimensions) is dominated by sites in Norrland. Skee is an outlier in the corner of the fourth quadrant (positive values on the first dimension and negative on the second). In the fourth quadrant other sites close to the Norwegian border are also found. The third dimension separates sites on the Finnish south coast (light) from the ones in Uppland (dark), and the Gotlandic (light) from the South Swedish ones (dark). Närpes has an extremely high value (white) in the third dimension.

Figure 3 displays the results of the MDS on a map using the RGB color model, which enables interpretation of the results even without any knowledge of the

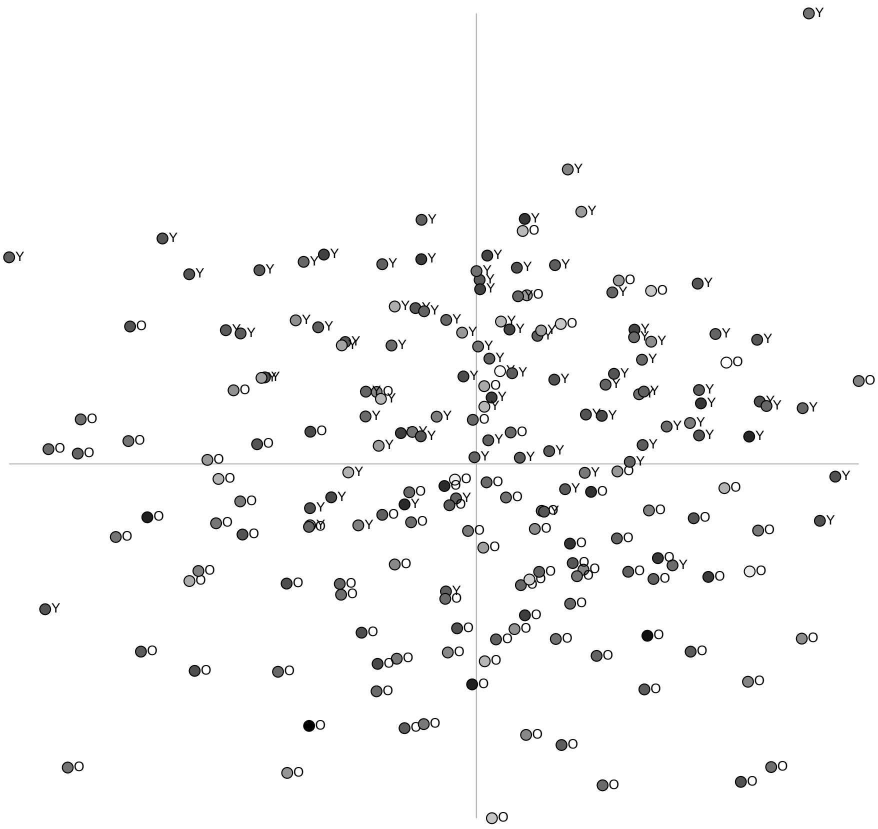


FIGURE 4: Results of MDS to 3 dimensions of the linguistic distances between sites and age groups. The 1st dimension is represented by the x-axis, the 2nd dimension by the y-axis and the 3rd dimension by the color of the dot. O = older speakers, Y = younger speakers.

geographic positions of the Swedish village names. In this map, the southernmost province, Skåne, forms a very coherent area with low values in the first and third dimensions and high values in the second dimension leading to green color. The separation of the South Swedish varieties from the ones on Gotland in the third dimension can be seen in colors close to cyan on Gotland. Uppland is also a very coherent area with orange color. Red and purple colors are found mostly close to the Norwegian border. In Norrland mostly dark green colors are found, but also other dark colors and blue. Götaland is quite incoherent with different colors from the center of the color spectrum. In Finland there is a clear difference between the sites on the south coast and the west coast.

The map shows that even if the distribution of dialectal features is continuous, some more coherent dialect areas can be detected.

[4.2] *Dialect leveling*

In the following step age-related variation was analyzed in addition to the geographic variation. The distance matrix that MDS was applied to comprised the pairwise linguistic distances between 196 objects (2 age groups × 98 sites). One dimension explains 78.9% of the variance, two dimensions 92.3% and three dimensions 95.9%.

Figure 4 on the preceding page shows the results of the MDS in a two-dimensional coordinate system where the color of the dot represents the third dimensions. The objects form one big cloud, except for one outlier, which for some reason has an extremely high value in the second dimension. This outlier is the younger speakers of Löderup (South Sweden). The labels of the objects do not fit into the plot, but for each object a letter indicates whether the dot concerns older or younger speakers. As can be seen, the second dimension mainly seems to separate the two age groups. The older speakers mostly have low values in the second dimension, while younger speakers have high values.

The maps in Figure 5 on the following page display the three dimensions of the MDS of older and younger speakers per site using the three-dimensional RGB color spectrum. The extremely high value of the younger speakers of Löderup in the second dimension would mean that a large proportion of the color representing the second dimension would be required for representing this variety. In order to produce more separation between the other varieties the young speakers of Löderup were left out of the color visualization. The two age groups are displayed on separate maps, but the colors of the two maps are comparable since they are based on one single MDS analysis.

The difference between the map of the older speakers and the map of the younger speakers is striking. In the map of the older speakers a broad spectrum of colors is found, while the map of the younger speakers is dominated by green. This shows a large-scale on-going leveling of the Swedish dialects. The dialect leveling can be confirmed statistically. The average linguistic distance between sites is larger for the older speakers than for younger speakers. This difference is statistically significant (Paired Samples t-test, $t(4752) = 30.1, p < 0.001$).

By comparing the colors of older and younger speakers in Figure 5 on the next page conclusions can be drawn about in which dialects vowel pronunciation is changing the most. For example, the sites in Finland have much more similar colors for older and younger speakers than many of the sites in Sweden. In order to get a more apparent view of which dialects that seem to be changing and which are stable, a map visualizing only the within site distances was created. The sites were divided into three groups using K-means clustering (the most commonly used clustering algorithm for partitioning data; the user decides how many groups should be formed, and the algorithm partitions the most similar items into groups by minimizing the total error sum of squares; Legendre &

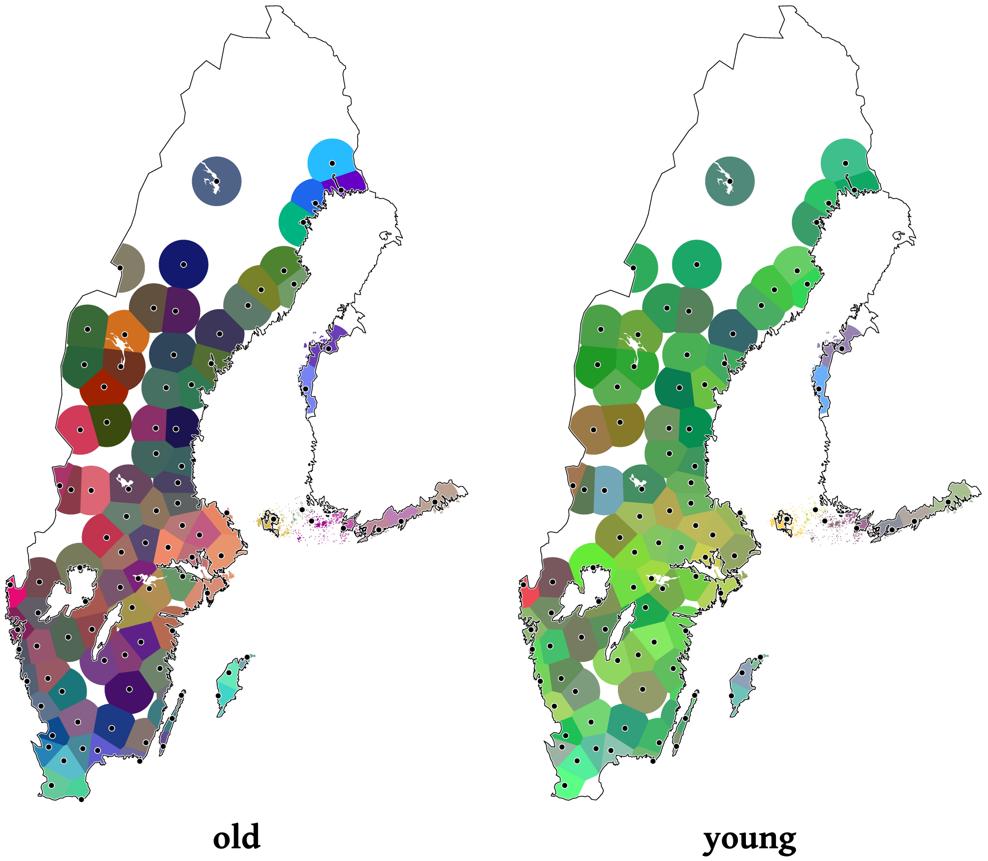


FIGURE 5: MDS to 3 dimensions of the linguistic distances between sites and age groups. Both age groups were included in one single MDS analysis, and are represented within the same color spectrum, making the colors of the two maps comparable with each other.

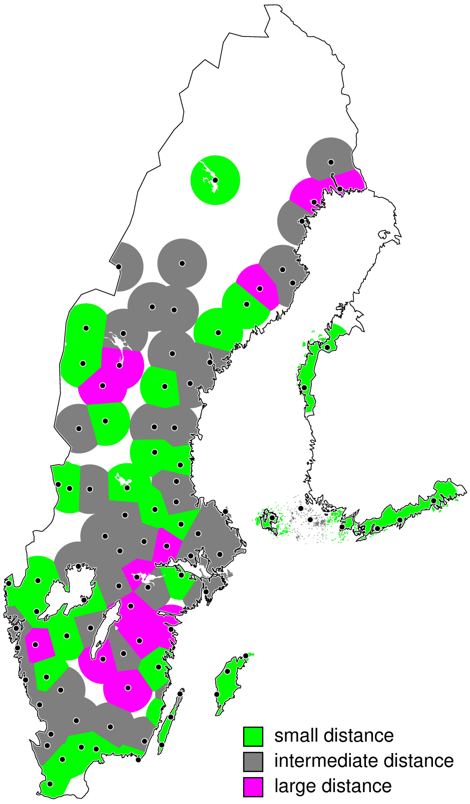


FIGURE 6: Map displaying the aggregate distances between older and younger speakers at each site. The sites were partitioned into three groups by K-means clustering.

Legendre 1998, 349–355). The map in Figure 6 shows the three groups obtained by clustering sites with the most similar distances between the two age groups. All sites with only a relatively small distance between older and younger speakers are green, sites with a large distance are magenta, and the sites with intermediate distances between older and younger speakers are gray.

Dialects in the South Swedish area, on the islands Öland and Gotland, and in Finland are green, and hence have small average distances in vowel pronunciation between older and younger speakers. The same holds for many sites around lake Vänern. These areas seem relatively stable when it comes to vowel pronunciation. Many of the sites close to the two biggest cities, Stockholm and Göteborg, and also in an area south-west from Stockholm are gray or magenta, which suggests a large ongoing change in vowel pronunciation. In Norrland there are sites of all three types: some dialects show a large ongoing change, some an intermediate change, and some are relatively stable.

In the map of the younger speakers in Figure 5 on the preceding page it looks as if there was almost no variation in vowel pronunciation between younger speakers. This is not entirely true. The variation between younger speakers is only so much smaller than between older speakers and between the two generations that only a small part of the color spectrum can be used for displaying the differences between younger speakers at different sites.

In order to be able to visualize dialectal differences within the younger age group, MDS was also applied separately to the older and the younger speakers. That is, two separate distance matrices were analyzed, one with the distances between older speakers at all sites and one with the distances between younger speakers at all sites. The analysis of only older speakers included 98 sites and the amount of explained variance was 82.1% for the first dimension, 93.4% for two dimensions and 95.7% for three dimensions. The analysis of only younger speakers included 97 sites (younger speakers in Löderup were left out) and the amount of explained variance was 83.8% for the first dimension, 92.4% for two dimensions and 96.1% for three dimensions.

Figure 7 on the facing page shows the maps with results of MDS applied separately to the older and younger speakers. Since two separate analyses are displayed in the two maps, the colors of the maps have to be interpreted independently. The maps are similar to the ones in Figure 5 on page 86, but the colors in each map are more distinct. When the whole color spectrum is used for each age group separately it becomes clear that there are differences across sites among the younger speakers, which could not be distinguished in Figure 5 on page 86. Moreover, the geographic pattern is quite similar for older and younger speakers. So even if the dialectal differences in vowel pronunciation are larger in the older generation than in the younger, the geographic distribution of dialectal features remains more or less constant.

Some differences in the geographic distributions can also be found. For example, the dialects in Norrland are more coherent in the younger age group than in the older. Figure 6 on the previous page showed that some dialects in Norrland are relatively stable, while others have a large linguistic distance between older and younger speakers. In Norrland the most divergent dialects seem to be changing the most, and thereby a more uniform spoken variety of Norrland is emerging. This can be seen as regionalization of the dialects, since the most divergent dialectal features seem to be disappearing while some other features that distinguish Norrland varieties from other Swedish varieties are preserved.

[5] DISCUSSION AND CONCLUSIONS

In this paper aggregating techniques were applied to Swedish vowel data in order to investigate geographic relationships in vowel pronunciation and language change in apparent time. The analyses showed that the variation in vowel pronunciation across Swedish dialects is continuous and no abrupt dialect borders exist. The absence of clearly separable dialect groups is in agreement with previous descriptions of the Swedish dialects. In the continuum a number of more

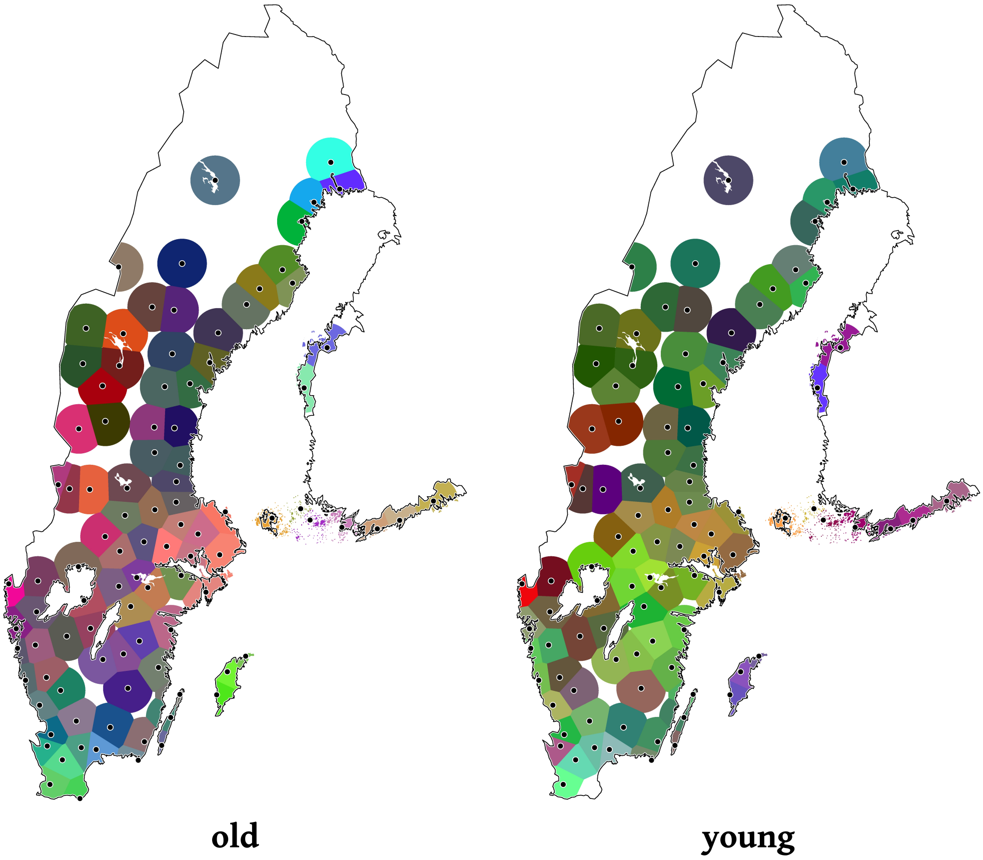


FIGURE 7: MDS to 3 dimensions of the linguistic distances between sites for each age group separately. The maps are based on two separate MDS analyses, so that the full color spectrum is used in each of the maps. The colors are not comparable across the two maps.

coherent dialect areas, separated by gradual transitions, can still be found. One should note that the statistical analyses of the data did not incorporate any geographic information, but purely linguistic data. Still, when interpreted in relation to geography the results show apparent geographical coherence, and the areas that can be detected in the maps displaying the results of multidimensional scaling agree to a large extent with the traditional division of Swedish dialects by Wessén (1969). The main dialect areas—South Swedish, Götaland, Svealand, Gotland, and Finland—can be distinguished in the continuum. Dialects close to the Norwegian border seem to share some features related to vowel pronunciation. The varieties in Finland fall into a Southern group, which shares features with the Svealand dialects, and a Western group. A similar split of the Finland-Swedish dialects was detected in an analysis of intonational variation by Bruce (2004).

The search for isogloss bundles in traditional dialectology has the restriction of being applicable only to geographic variation. One advantage of aggregation is that it "makes sense for many sorts of variation" (Nerbonne 2009). This was shown to be the case in the simultaneous analysis of geographic and age-related variation in Section [4.2] of this paper. The results of MDS show age-related variation mainly on the second dimension, but neither the age-related nor the geographic variation is abrupt but shows a continuous distribution. The analyses show a large-scale ongoing leveling of Swedish dialects. Leveling of the Swedish dialects has been observed by scholars during the latter half of the 20th century. Since leveling has been assumed to concern especially morphological, syntactical and lexical variation, while vowels are presumed to still show considerable dialectal variation, it is striking to find such strong evidence for dialect leveling taking only vowel pronunciation into account. The results also show the persistence of geographic regions. Even if the change in vowel pronunciation is large in some areas, the geographic distribution of dialectal features is not changing much, so that the main dialect areas remain the same even if the linguistic distances between varieties are getting smaller.

Nerbonne (2009) argues that the aggregation and abstraction over a large number of linguistic variables makes it possible to formulate more general characterizations of variation than the analysis of single features allows for. Aggregation gives a more reliable signal of provenance than single features do and makes it possible to identify dialect areas beyond bundles of isoglosses. Aggregation of large data sets also allows us to examine the question whether dialectal variation can best be described in terms of areas or continua. Aggregate analysis has successfully showed the global relationships in vowel pronunciation between the varieties analyzed in this paper. However, a drawback of aggregation is that while focusing on a general level of analysis, the linguistic structure of the variation is not revealed. Areas are detected and the amount of overall change can be measured, but questions like *how?* and *why?* are not answered. The aggregate analysis does not reveal which linguistic features characterize the dialect areas. For variationist linguists the reasons behind observed changes and geographic distribution patterns are essential since they can be related to historical developments, other expressions of human culture, social relationships and/or linguistic typologies.

In order to give an idea of the variation on the variable level in the present data set, Figure 8 on the following page displays one-standard-deviation ellipses of the 19 Swedish vowels of the older and younger speakers in the PC2/PC1 plane. The data for drawing the ellipses comprised the average PC values of both age groups at each site measured at the temporal midpoint of the vowel segments. By using average values per speaker group for drawing the ellipses, the individual variation within the groups has been filtered out, and the ellipses show the amount of linguistic variation across sites and across the two age groups. The graphs give an idea about the average position of each vowel in the PC space. The size and orientation of the ellipses indicate the amount of variation in each vowel and the main direction of the variation.

The general trend can be seen very clearly in Figure 8 on the next page: except from the leveling of dialects (smaller ellipses with less overlap for younger speakers than for older speakers) there is a general lowering of front vowels going on. Especially the long vowels in *dör*, *lär*, *lös*, *nät* and *söt* are being lowered by younger speakers and are thereby filling a place in the vowel space that was not previously filled by any Standard Swedish long vowel phoneme. Of the short vowels included in the data set the vowel in *lett* shows the most lowering. This lowering of front vowels in Swedish has first been described as a chain shift in the town Eskilstuna by Nordberg (1975). After that the lowering has been attested mainly in cities, for example by Hammermo (1989), Andersson (1994), Kotsinas (1994) and Aniansson (1996). Leinonen (2010) showed that the lowering of front mid-vowels has spread also to a large number of rural Swedish dialects. The areas where the vowel shift is strongest are areas close to and south-west of Stockholm and close to Göteborg, i.e. areas which show a large degree of change on an aggregate level in Figure 6 on page 87 in this paper.

The ongoing chain shift in Swedish front vowels is described in detail by Leinonen (2010). The chain shift can be regarded as simplification of the vowel system: the vowel inventory becomes smaller as allophonic variants of /ɛː/ and /øː/ are disappearing, and also more symmetrical filling up more of the potential vowel space as can be seen in Figure 8 on the next page. According to Kerswill (2002) simplification often occurs during regional dialect leveling. The leveling of Swedish dialects has been confirmed statistically by the aggregate analysis in this paper. The linguistic result of the leveling process, however, can be studied only by analysis on the variable level.

Areas that do not show much aggregate change in vowel pronunciation are

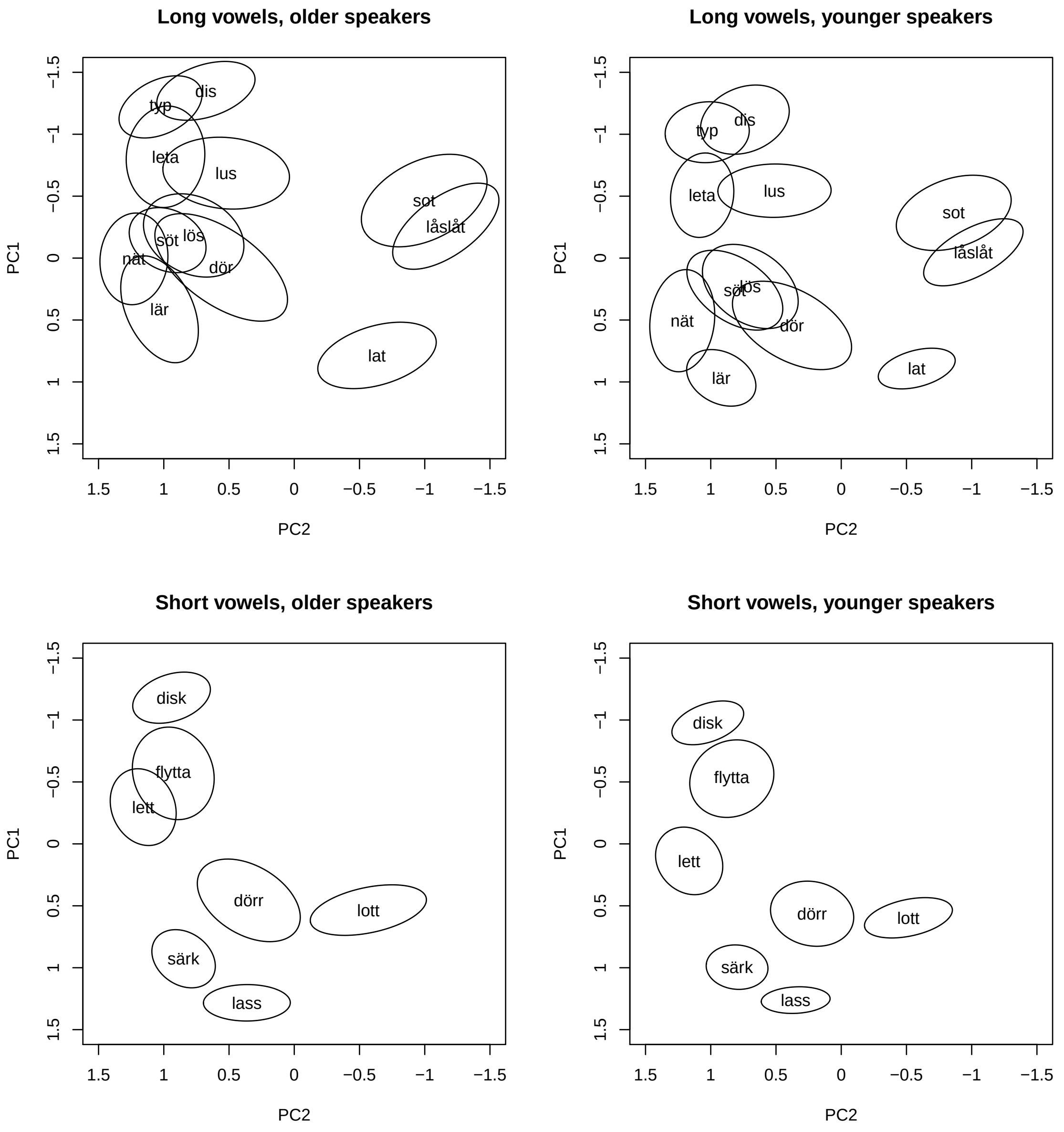


FIGURE 8: The 19 vowels of older and younger speakers in the PC2/PC1 plane. The one-standard-deviation ellipses are drawn based on the average PC values of the two speaker groups at each site measured at the temporal midpoint of each vowel.

South Sweden (mainly the province Skåne), Gotland and the Swedish dialect area in Finland. Edlund (2003, 28) has pointed out that Skåne and Gotland are Swedish regions with a strong regional identity. The identity is enhanced by an awareness of the historical developments that have formed these areas (Skåne was part of Denmark for a long time, while in medieval times Gotland was independent from Sweden and had an important position in the Hanseatic League), and by local traditions and cultural heritage. The Swedish language areas in Finland are separated from the rest of the Swedish dialects not only by the sea and a different political history, but also by a national border. Local identity is manifested through language use, and a strong identity serves to preserve characteristic features in the language.

Aggregation of linguistic data reveals global tendencies and overall relationships between varieties. It serves as a tool for detecting relationships which are not visible in an analysis of single features. For a more comprehensive understanding of linguistic variation, however, aggregate analysis should be complemented by detailed analysis of linguistic structure and extra-linguistic factors. Combining the results of different levels of aggregation and abstraction gives a better understanding of dialectal variation than when using the available methods separately.

REFERENCES

Andersson, Lars-Gunnar. 1994. Göteborgska – inte alltid så enkelt. In Ulla-Britt Kotsinas & John Helgander (eds.), *Dialektkontakt, språkkontakt och språkförändring i Norden*, vol. 40, MINS, 280–290. Stockholm: Institutionen för nordiska språk vid Stockholms universitet.

Aniansson, Eva. 1996. *Språklig och social identifikation hos barn i grundskoleåldern.* Ph.D. thesis, Uppsala University.

Bruce, Gösta. 2004. An intonational typology of Swedish. In *Proceedings of speech prosody 2004*, 175–178. Nara.

Bruce, Gösta. 2010. *Vår fonetiska geografi. Om svenskans accenter, melodi och uttal.* Lund: Studentlitteratur.

Edlund, Lars-Erik. 2003. Det svenska språklandskapet. De regionala språken och deras ställning idag – och i morgon. In Gunnstein Akselberg, Anne Marit Bødal & Helge Sandøy (eds.), *Nordisk dialektologi*, 11–49. Oslo: Novus.

Elert, Claes-Christian. 1994. Indelning och gränser inom området för den talade svenskan – en aktuell dialektografi. In Lars-Erik Edlund (ed.), *Kulturgränser – myt eller verklighet?*, vol. 4, DIABAS, 215–228. Institutionen för nordiska språk, Umeå universitet.

Elert, Claes-Christian. 2000. *Allmän och svensk fonetik*. Stockholm: Norstedt, 8th edn.

Engstrand, O., R. Bannert, G. Bruce, C.-C. Elert, O. Engstrand & A. Eriksson. 1997. Phonetics and phonology of Swedish dialects around the year 2000: A research plan. In *Rapporter från fonetik 97, den nionde svenska fonetikkonferensen*, vol. 4, PHONUM, 97–100. Umeå.

Eriksson, Anders. 2004. SweDia 2000: A Swedish dialect database. In P. J. Henrichsen (ed.), *Babylonian confusion resolved. Proceedings of the Nordic symposium on the comparison of languages*, vol. 1, Copenhagen Working Papers in LSP, 33–48.

Hallberg, Göran. 2005. Dialects and regional linguistic varieties in the 20th century I: Sweden and Finland. In Oskar Bandle (ed.), *The Nordic languages: An international handbook of the history of the North Germanic languages. 2*, vol. 22, Handbücher zur Sprach- und Kommunikationswissenschaft, chap. 185, 1691–1704. Berlin: de Gruyter.

Hammermo, Olle. 1989. *Språklig variation hos barn i grundskoleåldern*. Ph.D. thesis, Uppsala University.

Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, University of Groningen.

Jacobi, Irene. 2009. *On variation and change in diphthongs and long vowels of spoken Dutch*. Ph.D. thesis, University of Amsterdam.

Jain, Anil K. & Richard C. Dubes. 1988. *Algorithms for clustering data*. New Jersey: Prentice Hall.

Kerswill, Paul. 2002. Koineization and accommodation. In J. K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, 669–702. Malden, MA: Blackwell.

Kotsinas, Ulla-Britt. 1994. *Ungdomsspråk*, vol. 25, Ord och stil. Uppsala: Hallgren & Fallgren.

Legendre, Pierre & Louis Legendre. 1998. *Numerical ecology*, vol. 20, Developments in environmental modelling. Amsterdam: Elsevier, 2nd edn.

Leinonen, Therese. 2010. *An acoustic analysis of vowel pronunciation in Swedish dialects*. Ph.D. thesis, University of Groningen.

Lundberg, Jan. 2005. *Classifying dialects using cluster analysis*. Master's thesis, Göteborg University.

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1). 175–198.

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Edit distance and dialect proximity. In David Sankoff & Joseph Kruskal (eds.), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*, v–xv. Stanford: CSLI.

Nordberg, Bengt. 1975. Contemporary social variation as a stage in a long-term phonological change. In K.-H. Dahlstedt (ed.), *The Nordic languages and modern linguistics 2*, 587–608. Stockholm: Almqvist & Wiksell.

Peterson, Gordon E. & Harold L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24. 175–184.

Prokić, Jelena & John Nerbonne. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing* 2(1-2). 153–171.

Tabachnik, Barbara G. & Linda S. Fidell. 2007. *Using multivariate statistics*. Pearson, 5th edn.

Thelander, Mats. 2005. Sociolinguistic structures chronologically II: Swedish. In Oskar Bandle (ed.), *The Nordic languages: An international handbook of the history of the North Germanic languages. 2*, vol. 22, Handbücher zur Sprach- und Kommunikationswissenschaft, chap. 205, 1896–1907. Berlin: de Gruyter.

Tibshirani, Robert, Guenther Walter & Trevor Hastie. 2001. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society B* 63. 411–423.

Wessén, Elias. 1969. *Våra folkmål*. Stockholm: Fritzes, 9th edn. [first published 1935].

AUTHOR CONTACT INFORMATION

Therese Leinonen
Svenska litteratursällskapet i Finland r.f.
Riddaregatan 5
FIN-00170 Helsingfors
Finland
therese.leinonen@sls.fi

J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 97–112. (ISSN 1890-9639)
http://www.journals.uio.no/osla

COPING WITH VARIATION IN THE ICELANDIC PARSED HISTORICAL CORPUS (ICEPAHC)

EIRÍKUR RÖGNVALDSSON, ANTON KARL INGASON,
EINAR FREYR SIGURÐSSON
Reykjavík

ABSTRACT

We present an overview of an ongoing project which has the aim of developing methods for building a treebank of Icelandic. The treebank will contain texts from various different periods. Since Icelandic is an example of what has been called a less-resourced language when it comes to computational linguistics and language technology, it is essential to utilize the limited resources available as economically and efficiently as possible. We emphasize the importance of open source software and the interplay between linguistic knowledge and technological skills. We describe the workflow in the construction of the treebank and show how the different software tools work together towards the final representation. Finally, we show how the treebank can be used in studying some well known phenomena in Icelandic syntax.

[1] INTRODUCTION

In this paper, we describe an ongoing project, the Icelandic Parsed Historical Corpus (IcePaHC), which has the goal of developing economic and practical methods for building a treebank of Icelandic – methods which we hope can serve as a model in similar projects for other less-resourced languages. Icelandic is spoken by about 300,000 people and is clearly a less-resourced language (LRL) in any sense of the term. However, it has been the focus of much attention by syntacticians for the past two or three decades. There are several reasons for this. One is that due to its relatively rich morphology, Icelandic is ideal for testing several types of linguistic hypotheses. Another reason is that Icelandic has changed much less than its closest relatives and is thus ideal for testing and comparing theories of language change.

It is thus of great importance, not only to Icelandic syntacticians but to the general linguistic community, to have access to a large amount of well-structured data that enables researchers to study Icelandic syntax both synchronically and diachronically. As is well known, a syntactically parsed corpus – a treebank – is an important tool both for syntactic research and for the purposes of developing language technology tools. Our long-term goal is to build a treebank that will be useful both in syntactic research and for Icelandic language technology. The texts in our corpus will cover the history of Icelandic during a whole millennium, from the earliest written sources dating from the 12th century up to the present – approximately 100,000 words from each century.[1]

This paper describes our first steps in developing the treebanking methods and the treebank itself and is organized as follows. In section 2, we discuss briefly the motivations for building a parsed corpus like ours, and touch upon the challenges posed by the diversity of the texts. In section 3, we describe the software tools that we use and argue that an open source approach is essential for the development of NLP tools for less-resourced languages. In section 4, we describe the workflow in the construction of the treebank and show how the different software tools work together towards the final representation. Section 5 shows how the treebank can be used in the study of two well known phenomena in Icelandic syntax. Finally, section 6 is a conclusion.

[2] BACKGROUND AND CHALLENGES

Over the past two decades, interest in historical syntax has grown substantially among linguists. Accompanied by the growing amount of electronically available texts, this has led to the desire for – and possibility of – creating syntactically parsed corpora of historical texts, which could be used to facilitate search for examples of certain syntactic features and constructions. A few such corpora have been developed, the most notable being the Penn Parsed Corpora of Historical English, developed by Anthony Kroch and his associates (Kroch & Taylor 2000a; Kroch et al. 2004). These corpora have already proven their usefulness in a number of studies of older stages of English (cf., for instance, Kroch et al. 1995; Kroch & Taylor 2000b). We cooperate with the treebank team at the University of Pennsylvania and want to make our treebank compatible with their products – the Penn Treebank (Marcus et al. 1993) and the Penn Parsed Corpora of Historical English.

At a first glance, it may not seem feasible to build a diachronic treebank consisting of texts spanning a thousand years in the history of a language. However, Icelandic is often claimed to have undergone relatively small changes from the oldest written sources up to the present. The sound system, especially the vowel system, has changed dramatically, but these changes have not led to radical reduction or simplification of the system and hence they have not affected the

[1] At the time of the writing of this paper, a preview version (0.1) of the treebank (Wallenberg et al. 2010) which contains ca. 31,000 words from the 12th and 19th centuries has been released and can be downloaded from `http://www.linguist.is/icelandic_treebank/Download`. The corpus is released under the LGPL license which means that it may be freely distributed, modified and used in other software under certain restrictions - see `http://www.gnu.org/licenses/lgpl.html`. New versions will be released every three months until the project finishes, which is expected in mid-2011. We would like to thank Joel Wallenberg who has been instrumental in designing the corpus, Anthony Kroch and Beatrice Santorini for help and advice, and an anonymous reviewer for comments and suggestions.

inflectional system, which has not changed in any relevant respects. Thus, the morphosyntactic tagset developed for Modern Icelandic can be applied to earlier stages of the language without any modifications. The vocabulary has also been rather stable. Of course, a great number of new words (loanwords, derived words and compounds) have entered the language, but the majority of the Old Icelandic vocabulary is still in use in Modern Icelandic, even though many words are confined to more formal styles and may have an archaic flavor.

On the other hand, many features of the syntax have changed (cf. Faarlund 2004; Rögnvaldsson 2005). These changes involve for instance word order, especially within the verb phrase, the use of phonologically "empty" NPs in subject (and object) position, the introduction of the expletive *það* 'it, there', the development of new modal constructions such as *vera að* 'be in the process of' and *vera búinn að* 'have done/ finished', etc. The diversity of the texts obviously poses quite a challenge to the project. It is clear that both the methods of construction, the annotation scheme, the query language, and the search software will have to be able to deal with considerable variation in sentence structure.

If the goal of a project is to construct a parsed corpus of a less-resourced language like Icelandic, it is important to utilize whatever resources are available as efficiently as possible. Actually, one could argue that all languages other than English are, to varying degrees, less-resourced with respect to English. Thus the problems that the less-resourced language faces with respect to language technology are shared among the languages of the world. One can identify two main kinds of problems for languages other than English:

- The amount of people and money available to develop resources is small compared to what is available for English.

- The language is different from English in important linguistic ways and thus the established state of the art solutions need to be adapted from how they are applied to English.

Ideally, we would have liked to put together a group of experts, each of which has substantial cross-disciplinary knowledge about parsed corpora, artificial intelligence / machine learning, generative syntax and perhaps some more. In reality, we have a few people who specialize in some of those fields. This means that one of the keys to successful treebank construction for a language like Icelandic is defining interfaces between the technological knowledge and the linguistic knowledge. We discuss our approach to this problem in section [3.2]. The importance of being able to employ linguistic knowledge relates to the typological difference between English and the LRL because Icelandic, for example, differs from English in ways that are relevant to parsing. Icelandic has a rich morphology that affects any kind of an annotation process as opposed to English where issues of morphology can be ignored to a great extent. Another example is the V2 (verb second) constraint that has a substantial effect on Icelandic word order. Such a constraint does not apply to Modern English.

Related to those problems is the fact that the small number of speakers of a language like Icelandic means that there is limited interest in the field from the commercial sector. In order to attract such interest some measures must be taken to make existing resources as accessible as possible. In our case the most important measures of this kind are the releasing of a complete language processing toolkit under a free and open source license, as discussed in section [3.1].

[3] TECHNOLOGY AND LINGUISTICS AS RESOURCES

[3.1] *Open Source BLARK as a Foundation*

It has been noted that in order to do any kind of work on language technology for a given language a set of some basic tools, referred to as a BLARK (Basic Language Resource Kit, (cf. Krauwer 2003)), is the minimum requirement. A few such tools have been developed for Icelandic and packaged under the name IceNLP. Those include a rule based PoS (Part-of-Speech) tagger (Loftsson 2008), an HMM (Hidden Markov Model) tagger, a shallow parser (Loftsson & Rögnvaldsson 2007), a lemmatizer (Ingason et al. 2008), a sentence segmentizer and a tokenizer. The IceNLP toolkit has recently been made open source (LGPL-licensed) to encourage further innovation in Icelandic language technology.[2]

Open source licenses are important for language technology in general as researchers have pointed out (e.g. Halácsy et al. 2007; Forcada 2006). We believe that this importance is even greater in the context of an LRL such as Icelandic. An accessible BLARK without serious licensing barriers can make a difference for the LRL in two important ways:

- It attracts researchers and commercial innovators to work on language technology for the LRL.
- It encourages linking the LRL with other international open source projects.

Many language technology projects focus on developing so-called language independent solutions for various tasks. Despite being language independent in nature those efforts are somewhat limited by the fact that practical aspects of setting up experiments for many languages always take time and therefore evaluation of the methods in question rarely extends to a large number of languages. We believe that a complete open source package of basic tools for a language like Icelandic makes the language much more feasible for inclusion in such experiments. A researcher can download IceNLP and start tagging and lemmatizing

[2] IceNLP can be downloaded from: http://sourceforge.net/projects/icenlp/

Icelandic text in minutes without having to consider licensing restrictions. Since IceNLP is LGPL-licensed it is also feasible for commercial software developers to include its features as part of their products. An open source license also encourages linking the BLARK of the LRL with other international open source projects and in the case of IceNLP there are already a few ongoing projects of this sort that would not have been possible without an open source BLARK.[3]

[3.2] *Automated Corpus Revision using Linguistic Terminology*

The IceNLP package includes a format conversion utility named Formald. One of the features of this utility is the ability to get a labeled bracketing representation of the output. Although such a conversion does not contribute anything to the information structure by itself it allows us to further manipulate the data using tools that are designed for working with labeled bracketing. One such tool that has been very useful in our annotation process is CorpusSearch (CS) (Randall 2005).

As the name implies CS is a tool that can be used to search parsed corpora but it can also be used for automated rule-based corpus revision. The main strength of CS in this respect is the fact that it is designed for linguists and the query language allows the user to interact with a treebank using terminology that is familiar to a syntactician. Relationships are expressed using terms like *dominates, precedes, c-commands*, etc. This means that CS provides an abstraction layer between linguistics and technology. A person who is trained in syntactic theory can develop an advanced rule-based parser without knowing much about the technology that does the computational work behind the scenes. In our case such a parser is built on top of the output of IceNLP. While such an abstraction layer may not be the most theoretically interesting fact about the annotation process of a treebank it means a great deal in terms of getting practical results with limited resources.

Automated corpus revision in CS is based on revision queries like the one shown in (1).

(1) **A CorpusSearch revision query**

```
query: ({1}[1]NP* hasSister {2}[2]NP-POS)
       AND ([1]NP* iPrecedes [2]NP-POS)

extend_span{1, 2}:
```

The query means: If any kind of an NP (NP*) has a sister in the tree that is an NP-POS and the first NP immediately precedes the latter, the span of the first NP

[3] Those include context sensitive spelling correction for Icelandic based on LanguageTool (Naber 2003), a machine translation system based on Apertium (Forcada 2006) and a commercial project that involves automated market research. Discussion of those projects is beyond the scope of this paper.

should be extended so that it includes the latter.

The syntax of regular search queries is the same as for revision queries except they do not include revision commands such as *extend_span*. The non-technical syntactician can therefore also use this syntax to search for suspicious patterns that probably need manual correction. For example one could construct a query that searches for IPs that include more than one subject, again using linguistic terminology.

[4] BUILDING THE TREEBANK

[4.1] *IceNLP*

The workflow we use in the construction of the diachronic parsed corpus of Icelandic makes extensive use of IceNLP as well as other open source software. To illustrate this let us take a look at how one sentence is processed using IceNLP. The sentence in (2) is an example from Old Icelandic.

(2) Rannveig og Hergerður voru dætur þeirra
Rannveig and Hergerður were daughters their
'Rannveig and Hergerður were their daughters'

The first step in the automated annotation is to run the sentence through IceTagger to assign PoS-tags as exemplified in (3).

(3) **Output from IceTagger:**

```
Rannveig nven-m
og c
Hergerður nven-m
voru sfg3fþ
dætur nvfn
þeirra fphfe
```

Since the IceNLP tools use the Icelandic tags (cf. Loftsson 2008) we keep this representation for now but the tags are translated into English in a later step.

In the second step we use IceParser to perform shallow parsing (chunking of phrases) as shown in (4). In addition to marking phrases IceParser annotates some syntactic functions such as subjects and objects.

(4) **Output from IceParser:**

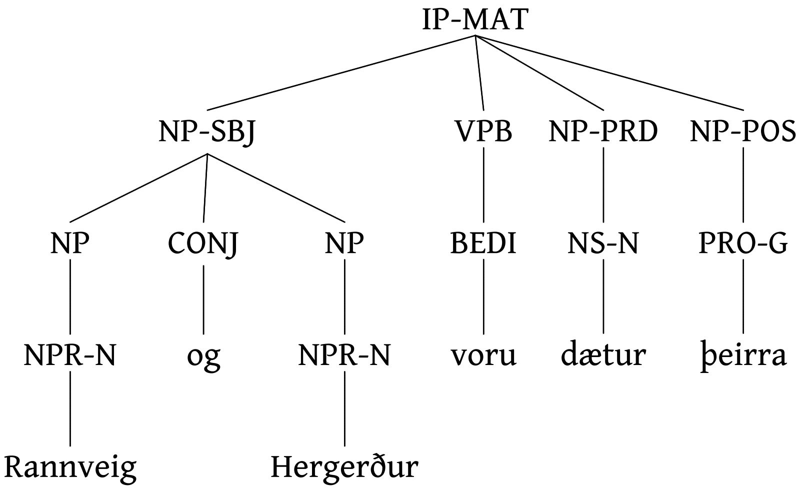
```
{*SUBJ> [NPs [NP Rannveig nven-m NP] [CP og c CP]
[NP Hergerður nven-m NP] NPs] *SUBJ>}
[VPb voru sfg3fþ VPB] {*COMP< [NP dætur nvfn NP] *COMP<}
{*QUAL [NP þeirra fphfe NP] *QUAL} . .
```

Then we use the lemmatizer Lemmald to assign a base form to each token in the sentence. The final step involving the IceNLP toolkit is to use one of its format conversion features to get a labeled bracketing representation of the sentence and translate the tagset to an annotation scheme that is mostly compatible with the Penn Corpora of Historical English. The result of those operations can be seen as labeled bracketing in (5) and as a tree diagram in (6). Note that lemmas are omitted from the tree diagrams in this paper.

(5) **Output after lemmatization and conversion to labeled bracketing:**

```
( (IP-MAT (NP-SBJ (NP (NPR-N Rannveig-rannveig) )
(CONJ og-og) (NP (NPR-N Hergerður-hergerður) ) )
(VPB (BEDI voru-vera) )
(NP-PRD (NS-N dætur-dóttir) )
(NP-POS (PRO-G þeirra-það) ) (. .-.) ) )
```

(6)



Thus, the diagram in (6) represents the kind of structure we can annotate using only the tools of the IceNLP toolkit.

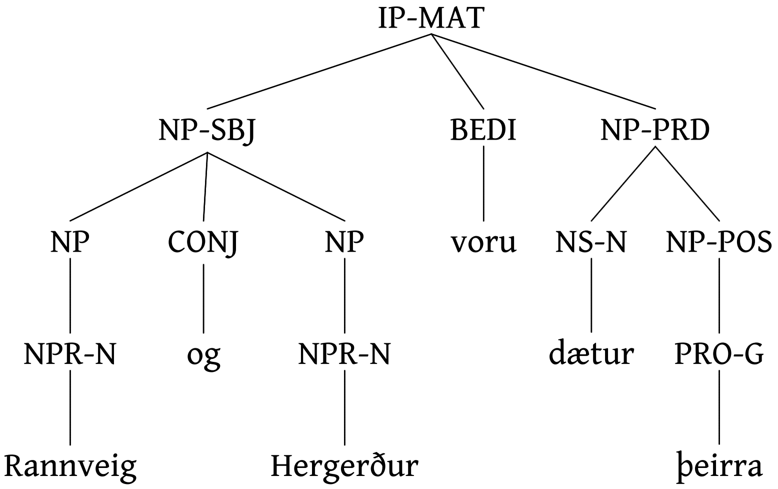
[4.2] *CorpusSearch and CorpusDraw*

The structure already contains a lot of information about the sentence but in order to finish the tree we use CS to apply revision queries to the structure. First we run the query in (1) so that NP-POS is moved under the immediately preceding NP. Finally we want the finite verb to be the head of the IP so we run the revision in (7) to delete VPs that are dominated by IPs and dominate finite verbs. Note that **finiteVerb** is defined by a regular expression that matches all finite verbs.

(7) **Revision query that removes extra VPs**

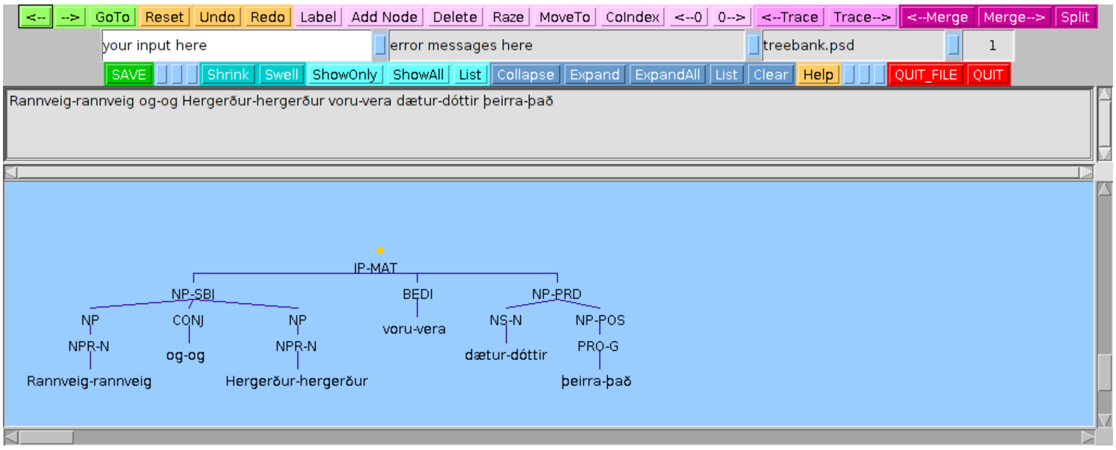
```
query: (IP-MAT iDoms {1}[1]VP*)
       AND ([1]VP* iDoms finiteVerb)

delete_node{1}:
```

(8)



The resulting tree is shown in (8). In this case, the automated rule-based annotation manages to generate the full structure we want. This is of course not always the case and while we aim to cover as many types of structure as possible automatically there are various examples of incomplete or wrong annotation that require manual corrections. Again, we find it important to design the workflow in a way that does not require a lot of technical knowledge, especially the parts that require extensive understanding of the theory of syntax (manual corrections occur at the linguistic end of the abstraction layer, not the technical one!). For this we use CorpusDraw, a program that is bundled with CS and provides a visual interface for correcting trees. A screenshot of the previous sentence from CorpusDraw is shown in (9).

(9) **CorpusDraw screenshot**


<-- --> GoTo Reset Undo Redo Label Add Node Delete Raze MoveTo CoIndex <--0 0--> <--Trace Trace--> <--Merge Merge--> Split
your input here
error messages here
treebank.psd
1
SAVE Shrink Swell ShowOnly ShowAll List Collapse Expand ExpandAll List Clear Help QUIT_FILE QUIT
IP-MAT
NP-SBJ
BEDI
NP-PRD
NP
CONJ
NP
NS-N
NP-POS
NPR-N
og-og
NPR-N
voru-vera
dætur-dóttir
PRO-G
Rannveig-rannveig
Hergerður-hergerður
þeirra-það

Rannveig-rannveig og-og Hergerður-hergerður voru-vera dætur-dóttir þeirra-það

[5] TWO CASE STUDIES

In this section, we show how the treebank could be used to study two phenomena in Icelandic syntax – DAT-NOM verbs and the so-called New Passive.

[5.1] *DAT-NOM verbs*

In Modern Icelandic (MIce) we get variation between number agreement and non-agreement with DAT-NOM verbs that take plural nominative objects, cf. (10-a) and (10-b).

(10)	a.	Stelpunni	lík**uðu**	strákar
		girl-THE-DAT	liked-3-**PLUR**	boys-NOM
		'The girl liked boys'		
	b.	Stelpunni	lík**aði**	strákar
		girl-THE-DAT	liked-3-**SING**	boys-NOM

This kind of variation is also found in Old Icelandic (OIce) (Eythórsson & Jónsson 2009), although nominative agreement seems to have been more frequent in OIce. Non-agreement, on the other hand, seems to be much more frequent in MIce than in OIce. That would indicate a change over the ages – a change we can study in IcePaHC. If there has been a significant change in the agreement system, that might explain why, for some speakers, DAT-ACC (for original DAT-NOM verbs, e.g. *líka* 'like') seems to be grammatical (Árnadóttir & Sigurðsson 2008).

We do a CorpusSearch query to see if a change has taken place. To get a clear idea of what we are dealing with, let us first look at the raw data we get from CorpusDraw for (10-a) above:

(11) **Raw data**

```
( (IP-MAT (NP-SBJ (PRO-D Stelpu$-stelpa) (D-D $nni-hinn))
          (VBDI líkuðu-líka)
          (NP-OB1 (NS-N strákar-strákur))
          (. .-.)))
```

Now we can define the search which finds agreement as well as non-agreement with plural objects of DAT-NOM verbs, cf. (12).

(12) **A CorpusSearch query for DAT-NOM verbs**

```
node: IP*
 query: (IP-MAT*|IP-SUB* iDoms NP-SBJ)
        AND (IP-MAT*|IP-SUB* iDoms NP-OB1)
        AND (IP-MAT*|IP-SUB* iDoms !VAN*)
        AND (NP-SBJ iDoms *-D)
        AND (NP-OB1 iDoms NS-N)
```

The query matches any main clause (IP-MAT*) or embedded clause (IP-SUB*) that immediately dominates (iDoms) a subject (NP-SBJ) and an object (NP-OB1) and which does not immediately dominate a passive participle (!VAN*) (the '!' negates the matched element) since we do not want to include passives in our results. Furthermore, the subject phrase immediately dominates a nominal element in the dative case (*-D), where the star is a wildcard that matches nouns, pronouns, determiners and quantifiers. The object phrase immediately dominates a plural nominative noun (NS-N).

After running the CS query, we are able to compare relative frequencies of agreement vs. non-agreement from different periods of the history of Icelandic.

Since the treebank is compatible with the Penn Parsed Corpora of Historical English same, or similar, phenomena in Icelandic and English at various stages can be compared. Let us, for example, take a look at the raw data for the following sentence from Early Modern English (Kroch et al. 2004):

(13) **Early Modern English raw data**

```
( (IP-MAT (NP-SBJ (PRO I))
          (VBP believe)
          (CP-THT (C 0)
                  (IP-SUB (NP-SBJ (PRO I))
                  (MD shall)
                  (VB like)
                  (NP-OB1 (PRO$ your) (N cook))
                  (ADVP (ADV very) (ADV well))))
          (. .)) (ID FHATTON-E3-H,I,148.34))
```

In this example we have the main verb *like* in the embedded clause (IP-SUB). As can be seen, the parsing and the labels are (almost) the same as in IcePaHC. That makes it a lot easier to do a comparative study of those languages.

[5.2] *The New Passive*

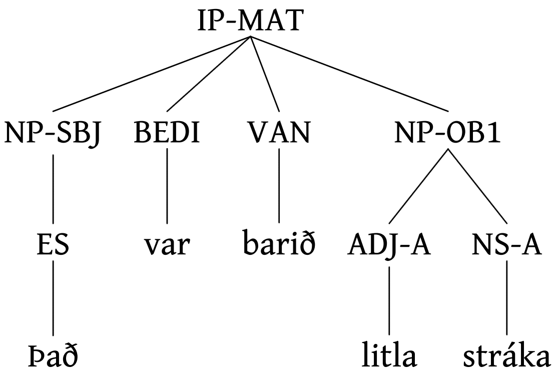
There has been a lively discussion about the New Passive in Icelandic in the recent years and opinions differ widely on its nature. Some researchers claim it is not a passive at all, but instead an active (cf. Maling & Sigurjónsdottir 2002), whereas others claim that it is simply a new form of the passive (e.g. Eythórsson 2008). Although it uses passive morphology, the object always stays in situ and does not undergo NP-movement (A-movement), cf. (15). Furthermore, it can be in the accusative case, and does not trigger agreement of the finite verb and the participle as a nominative argument in the canonical passive would do. Instead, the finite verb is always 3sg and the passive participle (the main verb), which assigns case (contra Burzio's Generalization), is neuter singular. (14) shows different versions of the canonical passive, whereas (15) shows the New Passive.

(14) a. Það voru barðir litlir
it were beaten-MASC-PLUR little-MASC-NOM-PLUR
strákar
boys-MASC-NOM
'Little boys were beaten'
b. Það voru litlir strákar barðir
c. Litlir strákar voru barðir

(15) a. Það var barið litla stráka
it was beaten-NEUT-SING little-MASC-ACC-PLUR boys-MASC-ACC
'Little boys were beaten'
b. *Það var lítinn strák barið
it was little-MASC-ACC-SING boy-MASC-ACC beaten-NEUT-SING
'A little boy was beaten' (Eythórsson 2008, 213, ex. (76))

In the tree diagram in (16) we show how (15-a) would be parsed in our corpus. Notice that the expletive *það* is tagged ES.

(16)



As in the canonical passive, the verb *vera* 'be' (or *verða* 'will be, become') is always used in the New Passive. As expected, it is not always finite.

(17) Það hefur oft verið barið litla
it has often been beaten-NEUT-SING little-MASC-ACC-PLUR
stráka
boys-MASC-ACC
'Little boys have often been beaten'

Even though the New Passive sentences begin with the expletive *það* in the examples above, this is not always so, as seen in (18) and (19). Furthermore, as shown in (19), the passive participle does not always follow *vera*/*verða*. It can precede the verb in sentences where Stylistic Fronting has applied. In the absence of an overt expletive *það* we include an empty category **exp** in the annotation.

(18) Í gær var barið litla
yesterday was beaten-NEUT-SING little-MASC-ACC-PLUR
stráka
boys-MASC-ACC
'Little boys were beaten yesterday'

(19) Skoðað verður miða við innganginn
inspected-NEUT-SING will.be tickets-MASC-ACC on entrance-THE
'Tickets will be inspected on entering' (Maling 2006, 200, ex. (7))

Example (19) above shows that we cannot rely on the main verb immediately preceding the object.

From the facts described above we can use the following for a New Passive search query (with accusative object):

(20) a. It contains an expletive (overt or covert)
b. It contains the verb *vera* 'be' (BE*) or *verða* 'will be, become' (RD*)
c. It contains a passive participle (tagged as VAN)
d. It contains an object (NP-OB1)
e. The direct object is in accusative case

Following these facts (in that order), the CS query might look like this:

(21) **A CorpusSearch query for the New Passive**

```
node: IP*
query: (IP* idoms NP-SBJ)
       AND (NP-SBJ idoms ES|\*exp\*)
       AND (IP* iDoms BE*|RD*)
       AND (IP* iDoms VAN)
       AND (VAN hasSister NP-OB1)
       AND (NP-OB1 iDoms *-A)
```

Even though the literature on the innovative New Passive – which is almost exclusively found in texts from the late 20th century up to the present day – is already quite extensive, many things regarding its nature and origin remain unclear and disputed. The question arises, of course, why a 20th century child would re-analyse passive sentences. In other words, what is the source of the New Passive? Various attempts – which will not be repeated here – have been made to answer the question.

One possible factor could be that the use of the expletive *það* increased heavily in the early 19th century as Hróarsdóttir (1998) shows (cf. also Rögnvaldsson 2002). This can lead to the subject of the canonical passive not being A-moved, as shown in (14a). These possible effects on the New Passive cannot be fully investigated without a diachronic treebank.

[6] CONCLUSION

In this paper we have presented the outlines of our work in developing efficient methods for building a treebank of a less resourced language – Icelandic in our case. This is still very much a work in progress but we think that our approach could serve as an example for other less-resourced languages. We have emphasized the re-use of existing tools and the importance of open source policy in this respect. We have also emphasized the importance of linguistic insights and the interplay between linguistic knowledge and technological skills in developing software tools for building syntactic trees. We described the workflow in the construction of IcePaHC and presented examples of how it can be used to study celebrated constructions in Icelandic.

Obviously, we are far from having a full-fledged treebank at our disposal. However, we feel that we have come quite far in developing the methods for building the treebank, and we have already started the actual production of trees.

REFERENCES

Árnadóttir, H. & E. F. Sigurðsson. 2008. The glory of non-agreement: The rise of a new passive. Ms., University of Iceland.

Eythórsson, T. 2008. The New Passive in Icelandic really is a passive. In T. Eythórsson (ed.), *Grammatical Change and Linguistic Theory. The Rosendal papers*, 173–219. Amsterdam: John Benjamins.

Eythórsson, T. & J. G. Jónsson. 2009. Variation in Icelandic morphosyntax. In A. Dufter, J. Fleischer & G. Seiler (eds.), *Describing and Modeling Variation in Grammar*, Trends in Linguistics. Studies and Monographs 204, 81–96. Berlin: Mouton de Gruyter.

Faarlund, J. T. 2004. *The Syntax of Old Norse*. Oxford University Press.

Forcada, M. L. 2006. Open-Source Machine Translation: an Opportunity for Minor Languages. In *Strategies for Developing Machine Translation for Minority Languages (5th SALTMIL Workshop on Minority Languages) (organized in conjunction with LREC 2006)*, 8–15. Genova.

Halácsy, P., A. Kornai & C. Oravecz. 2007. Hunpos – an open source trigram tagger. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, 209–212.

Hróarsdóttir, T. 1998. *Setningafræðilegar breytingar á 19. öld. Þróun þriggja málbreytinga*. Reykjavík: Institute of Linguistics, University of Iceland.

Ingason, A. K., S. Helgadóttir, H. Loftsson & E. Rögnvaldsson. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In *GoTAL '08: Proceedings of the 6th international conference on Advances in Natural Language Processing*, 205–216. Berlin, Heidelberg: Springer-Verlag. doi: http://dx.doi.org/10.1007/978-3-540-85287-2\\\\\\\_20.

Krauwer, S. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*, 8–15.

Kroch, A., B. Santorini & L. Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM. First Edition. Size: 1.8 million words.

Kroch, A. & A. Taylor. 2000a. Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second Edition. Size: 1.3 million words.

Kroch, A. & A. Taylor. 2000b. Verb-Object Order in Early Middle English. In *Diachronic Syntax: Models and Mechanisms*, 132–163. Oxford University Press.

Kroch, A. S., A. Taylor & D. Ringe. 1995. The Middle English verb-second constraint: a case study in language contact and language change. In Susan Herring et al (ed.), *Textual Parameters in Older Language*. Amsterdam: John Benjamins.

Loftsson, H. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics* 31(1). 47–72.

Loftsson, H. & E. Rögnvaldsson. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the* 16^{th} *Nordic Conference of Computational Linguistics (NoDaLiDa 2007)*. Tartu, Estonia.

Maling, J. 2006. From passive to active. Syntactic change in progress in Icelandic. In B. Lyngfelt & T. Solstad (eds.), *Demoting the Agent. Passive, middle and other voice phenomena*, 197–223. Amsterdam: John Benjamins.

Maling, J. & S. Sigurjónsdóttir. 2002. The 'new impersonal' construction in Icelandic. *Journal of Comparative Germanic Linguistics* 5. 97–142.

Marcus, M. P., B. Santorini & M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330.

Naber, D. 2003. *A Rule-Based Style and Grammar Checker*. Diploma thesis, University of Bielefeld. URL http://www.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf.

Randall, B. 2005. *Corpussearch 2 user's guide*. University of Pennsylvania. URL http://corpussearch.sourceforge.net/CS-manual/Contents.html.

Rögnvaldsson, E. 2002. ÞAÐ í fornu máli – og síðar. *Íslenskt mál* 24. 7–30.

Rögnvaldsson, E. 2005. Setningafræðilegar breytingar í íslensku. In H. Thráinsson (ed.), *Setningar. Handbók um setningafræði. Íslensk tunga 3*, 602–635. Reykjavík: Almenna bókafélagið.

Wallenberg, J. C., A. K. Ingason, E. F. Sigurðsson & E. Rögnvaldsson. 2010. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.1. Size: 31 thousand words. URL http://www.linguist.is/icelandic_treebank.

AUTHOR CONTACT INFORMATION

Eiríkur Rögnvaldsson
University of Iceland
Dept. of Icelandic
IS-101 Reykjavík
Iceland
eirikur@hi.is

Anton Karl Ingason
University of Iceland
Dept. of Icelandic
IS-101 Reykjavík
Iceland
anton.karl.ingason@gmail.com

Einar Freyr Sigurðsson
University of Iceland
Dept. of Icelandic
IS-101 Reykjavík
Iceland
einarfs@gmail.com

OSLa J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 113–128. (ISSN 1890-9639)
http://www.journals.uio.no/osla

LINGUATECA'S INFRASTRUCTURE FOR PORTUGUESE AND HOW IT ALLOWS THE DETAILED STUDY OF LANGUAGE VARIETIES

DIANA SANTOS
SINTEF, Oslo & FCCN, Lisboa

ABSTRACT

In this paper I present briefly Linguateca, an infrastructure project for Portuguese which is over ten years old, showing how it provides several possibilities to study grammatical and semantic differences between varieties of the language.

After a short history of Portuguese corpus linguistics, presenting the main projects in the area, I discuss in some detail the AC/DC project[1] and what is called the AC/DC cluster (encompassing other related corpus projects sharing the same core). Emphasizing its potential for language variation studies, the paper also (i) describes CONDIVport's integration as an impetus for new capabilities, and (ii) provides a sketch of newly added functionalities to AC/DC.

[1] AN INFRASTRUCTURE FOR PORTUGUESE LANGUAGE TECHNOLOGY

In line with the main audience of the Linguateca project, there have already been several descriptions of Linguateca as an infra-structure for Portuguese, in Portuguese (Santos 2009), as well as substantial reporting[2]. However, an international audience has only seen traces and scattered references so far, so this paper intends to fill this gap in what corpus resources are concerned.

It all started in 1998, with a small project (1998-1999) for preparing the future of the computational processing of the Portuguese language hosted by SINTEF, as an area to be taken specially good care of in the future science and technology programme (for the *White book on Science and Technology* created by the then Ministry of Science and Technology in Portugal), and wrote a small memo to be discussed publicly by all interested parties (Santos 1999a).

[1] The name stands for *Acesso a Corpos, Disponibilização de Corpos* (roughly: access to corpora, making corpora available), and is meant to signal that it should both benefit users – granting them access; and corpus owners: helping them to make their corpora widely available. See www.linguateca.pt/ACDC/

[2] More than 500 items in the publication list at http://www.linguateca.pt

After the discussion, and given that several projects had been started (catalogue, publications catalogue, and some corpus services), one more year was granted, that prepared the ground for what later became *Linguateca.*

Linguateca was conceived as a three-axed initiative to foster R&D in the computational processing of the Portuguese language, with relevant work on (i) information dissemination, (ii) resource creation, and (iii) organization of evaluation initiatives.

[2] PORTUGUESE TEXT CORPORA

Assuming that an international audience is probably generally unaware of what has been done in Portuguese corpus processing, I will attempt here a short presentation of the field, with special emphasis on what is offered by Linguateca.

[2.1] *A brief history*

As far as I know, corpus compilation for Portuguese started during the 1960s with the *Português Fundamental* (Bacelar do Nascimento et al. 1984, 1987), a project shaped after and inspired by the *Français Fondamental* (Gougenheim et al. 1964). Strict criteria for documenting authentic usage in oral contexts all over the country were used, and a significant number of documents of spoken Portuguese (from 1971 to 1974) was recorded, transcribed and analysed at the Centro de Linguística da Universidade de Lisboa, see Bacelar do Nascimento (2001). The work of this team has continued ever since with the compilation i.a. of the large *Corpus de Referência do Português Contemporâneo*, CRPC[3] (Bacelar do Nascimento 2000), as can also be appreciated in the recent papers on the comparison of African varieties of Portuguese (Bacelar do Nascimento et al. 2008a,b).

Several degrees of latitude and longitude further, the NURC project (Callou 1999) was taking place in Brazil, aiming to describe the oral and educated language[4] in five major Brazilian cities (Recife, Salvador, Rio de Janeiro, São Paulo and Porto Alegre), being thus a five-headed project. Started in 1970, it produced different oral corpora and different research lines, as can be better appreciated in the overview by Varejão (2009). In NURC-RJ, comparative oral corpora of the decades 1970s and 1990s were deployed, and it is currently connected with the project *Para uma História do Português do Brasil*, PHPB[5], including also written materials since the XVIth century. In Recife, the project was extended to address conversation analysis, while in Porto Alegre it merged with the VARSUL project (Menon et al. 2009).

Outside a Portuguese-speaking countries context, Brigham Young University

[3] "Reference Corpus of contemporary Portuguese", see http://www.clul.ul.pt/sectores/linguistica_de_corpus/projecto_crpc.php

[4] Norma URbana Culta, see e.g.http://www.letras.ufrj.br/nurc-rj/

[5] "for a history of Brazilian Portuguese", http://www.letras.ufrj.br/phpb-rj/

(US) researchers were interested in electronically available Portuguese material, having created the Borba-Ramsey corpus[6], a subset of which was later included in the European Corpus Initiative (Thomson et al. 1994) and has since 1999 been browsable also through AC/DC. We can also mention Portext (Maciel 1997) in France, the English-Norwegian Parallel Corpus in Norway (Oksefjell 1999) and the VISL (Bick 1997) project in Denmark, as early providers of Portuguese texts searchable on the web. Castilho et al. (1995) mention John Uriagereka from Maryland as having proposed a joint database for Portuguese and Gallician as early as 1991. From the same source we also learn that in 1993 there was already a corpus project in Mozambique, led by Perpétua Gonçalves.

As to the specific comparison of different varieties of Portuguese, there are at least six corpus-based projects that deserve mention here: The Tycho Brahe project (Galves 2009), VARPORT (Brandão & Mota 2003), PEPB (Peres & Kato 2004), Corpus do Português (Davies & Ferreira 2006-), Banco do Português (Berber Sardinha 2007), and CONDIVport (Soares da Silva 2010). Early corpus-based work can be found in Barreiro et al. (1996); Wittmann et al. (1995).

For further information and historical overviews on Portuguese corpora – of which the pointers presented are just a small part, since many other corpora have come to light during the last decade – see Bacelar do Nascimento et al. (1996), Oksefjell & Santos (1998), Berber Sardinha (1999), Davies (2008), Berber Sardinha & Almeida (2008), Santos (2009) and Varejão (2009), as well as, of course, Linguateca's resource catalogue.

What I would like to stress here, before introducing the AC/DC project in the next section, is: when it started back in 1998, there were no services on the web that allowed a linguist or an engineer to query a Portuguese corpus. Also, the few available corpora for download had very different formatting, encoding, and conceptual organization, so that their content was hard to compare and required a lot of processing to be used simultaneously, as explained in Santos (1999b) as initial motivation for AC/DC.

[2.2] *The AC/DC cluster*

As devised in 1998-1999, AC/DC had as its main purpose to make a large number of corpus resources available on the web with a unified and simple interface that allowed people to interact with corpora without requiring physical access to institutions or software installation (at that time, there was no such thing for Portuguese). Later on we also considered as Linguateca's task to create resources that were lacking, such as a large newpaper text corpus, CETEMPúblico (Santos & Rocha 2001), which was also included in the AC/DC service.

As a service to the (Portuguese-language processing) community, every corpus owner or developer could make use of AC/DC to serve his corpora, and we

[6] Named after the corpus compilers, Francisco Borba and Myriam Ramsey.

have in fact tried to contact everyone and make the offer explicit, for modern Portuguese. In some cases, however, the offer was turned down (or simply ignored), for reasons that ranged from copyright problems to the desire of the particular groups to develop their own solutions. We note, however, that no requirement of exclusivity was ever made by Linguateca: on the contrary, our own corpora, notably CETEMPúblico, were also distributed by the Linguistic Data Consortium (LDC) and by Mark Davies for some time. So, one of the most used corpus of Portuguese, the NILC corpus, was given access by AC/DC although many other solutions to make it available were created as well by NILC (Aluísio et al. 2004).

Other related (resource) services provided by Linguateca were then developed as, in a way, an outgrowth of the basic AC/DC services, and I refer to this extended set as the AC/DC cluster, including the Floresta Sintáctica treebank (Afonso et al. 2002; Freitas et al. 2008) – the first treebank for Portuguese, COMPARA (Frankenberg-Garcia & Santos 2003) – a large manually revised Portuguese-English fiction parallel corpus, and CorTrad (Tagnin et al. 2009) – a parallel (multi-version and multi-genre) corpus. These other resources have further tools, parts, and interfaces, which will not be dealt with here, and were created in cooperation with other researchers and projects.[7]

[3] STUDYING VARIATION AND LANGUAGE VARIETIES WITH THE AC/DC CLUSTER

I start by a presentation of the available material, then present the browsing of CONDIVport, which was compiled for variational analysis, and finally present new functionalities for corpus-based discovery that are currently under test in the AC/DC project.

[3.1] *The initial and obvious data gathering*

In order to be able to compare and study varieties and variation, one has to have materials that represent them. So, the first and obvious requirement is to have plenty of material, so that one can take a "language bath", and immerse in language use in different countries, times and social classes. While this seems easy and obvious, in practice it isn't. In fact, what most people have in terms of electronic corpora is opportunistically gathered in nature, and Linguateca's offer is no exception.

In Table 1 on the facing page, the AC/DC material is roughly quantified under the genre parameter. Of course genre is a very elusive category, and a really thorough study of Portuguese genre is still unavailable, so under "informative, technical" different subcategories were joined such as essay, encyclopedic and textbook material, as well as email on librarianship. Also, thematic newspaper

[7] See `http://www.linguateca.pt/Floresta/`, `http://www.linguateca.pt/COMPARA/`, and `http://www.fflch.usp.br/dlm/comet/consulta_cortrad.html` for more information.

TABLE 1: Genre distribution in the AC/DC cluster, as of July 2010

Genre	Size in words
Narrative fiction	17,208,056
General newspaper	246,112,499
Specialized newspaper	6,367,807
Informative, technical	4,489,043
Oral	500,811
Other or not classified	5,067,371

corpora were classified as "specialized newspaper", while local and (party) political newspapers were considered "general newspaper". As to the "other" category, it joins at least (e-mail) spam, EU calls, business letters, legal documents and web texts, especially blogs.

Table 2 presents the material in terms of language variety.

TABLE 2: Variety distribution in the AC/DC cluster, as of July 2010

Language variety	Size in words
Africa	76,802
Brazil	64,878,821
Portugal	215,377,125
Unknown	723,626

Finally, just for the fiction material, Table 3 on the following page presents the distribution per decade in the last two centuries. Since three of the sources concern parallel corpora, let me clarify that only the material in Portuguese is counted (and the dates for the translation concern the publication of the translation, not of the original). For more details see the corresponding project pages. Note also that literary text of which the exact sources are not known (such as those included in some multi-genre corpora in AC/DC) is not included.

In addition to the textual material, special sentence separation and tokenizer modules for Portuguese were developed in AC/DC, and all data were parsed by PALAVRAS (Bick 2000), offering lemma, part of speech, morphological information (such as tense form, gender, number, pronoun case, diminutive, aumentative and superlative degree) and syntactical function (in a version of dependency structure constraint grammar developed for Portuguese by Eckhard Bick, including also some discourse-related features such as topic and focus and some semantic information). As discussed in Inácio & Santos (2006), some of the material in the AC/DC cluster has been manually revised, as to their text and to their annotation, but most of it has not (after all, AC/DC encompasses more than 280 million

TABLE 3: Temporal distribution of literary texts in the AC/DC cluster (tokens) (July 2010)

Decade	Vercial	COMPARA	ENPC	CorTrad
1800	207,473			
1810	252,599			
1820	229,116			
1830	53,110			
1840	323,622			
1850	89,258	11,302		
1860	591,702	22,053		
1870	511,453	18,766		
1880	666,540	84,549		
1890	304,846	17,055		
1900	543,050	29,937		
1910	377,369	21,840		
1920	328,588	5,943		
1930	103,136			
1940				
1950				
1960		17,802		
1970		160,240		
1980		256,423		
1990		764,942	72,389	
2000		25,074		98,806

words, or ca.16 million different sentences).

In addition to having developed our own AC/DC format as a transduction of PALAVRAS output format, we have also started to add semantic information in some domains, using a simple lexicon-driven approach followed by human rule writing for correction and improvement of both precision and recall, as described in Silva & Santos (2009); Santos & Mota (2010).

The distribution of the colour domain can be appreciated in Figure 1 on the next page, where both the density of colour tokens and types is shown. As far as I know, this is the largest semantically annotated corpus, which has undergone human revision, currently available. (Although colour annotation of the largest corpora has not yet been fully revised.)

[3.2] *Support for formal variational linguistics*

In addition to providing an "electronic bookshelf", or a web distribution window, to any group or project that is willing to have us making their corpus or resource

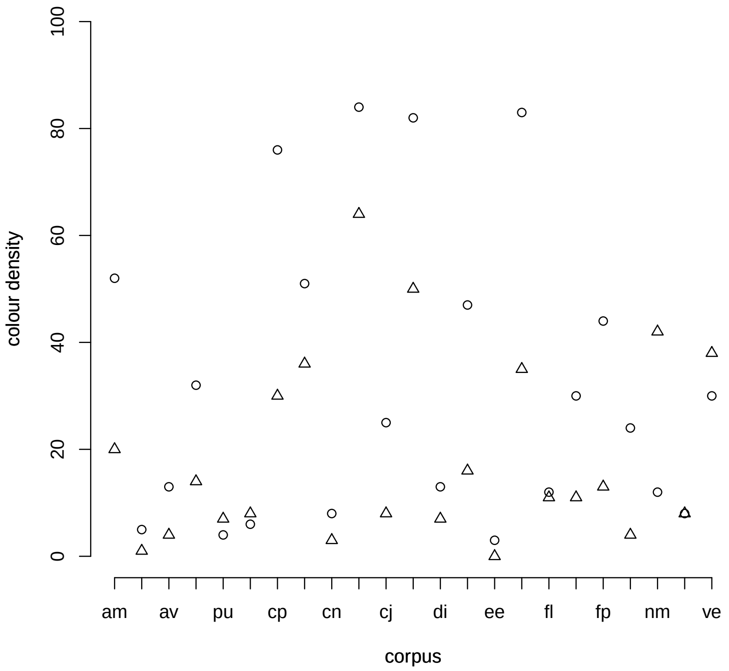


FIGURE 1: The semantic field of colour in AC/DC as of July 2010: circles describe types, triangles tokens. Colour density is defined as 10,000 times the ratio of colour words (tokens or types) compared to all words (tokens or types) in the corpus.

available, the AC/DC project may also develop specific facilities for the resources it (re)distributes, if they display new features.

This happened with the CONDIVport corpus – we started by simply making it available through the web as a regular AC/DC member, but soon we understood the interest in providing support for more complex models of (on-line) linguistic research: Given that CONDIVport was compiled to study the convergence and divergence of national varieties of Portuguese, under the framework initially developed by the *Quantitative Lexicology and Variational Linguistics group* in Leuwen[8], it had, in addition to three specific themes (soccer, fashion and health), texts from three different time periods, from Brazil and from Portugal. In addition, as an integral part of the methodology, a list of terms in the two first of these themes had also been compiled.

For foundations and critical discussions of the methodology, I redirect the reader to Geeraerts et al. (1999); Geeraerts & Grondelaers (1999); Speelman et al. (2003); Soares da Silva (2010). Here, I will only provide concrete examples of how

[8] See http://wwwling.arts.kuleuven.ac.be/qlvl/

the process goes: First, one gathers a set of *formal onomasiological profiles*[9] for key concepts in a given area – let us take clothing as an example: key concepts may be BLUSA (roughly “blouse”) or SAIA (roughly “skirt”). Their onomasiological profile is a set of lexical items which the linguist classifies (in context) as belonging to this class. So, as an example, the CASACO F (“female overcoat”) profile has been found to be: *blazer*, *blêizer*, *casaco*, *casaquinho*, *casaquinha*, *manteau*, *mantô*, *paletó*, *paletot* (Soares da Silva 2008a, page 66).

Together with their frequencies, these profiles allow the researcher to compute several measures such as uniformity, and relative uniformity – an example from soccer (Soares da Silva 2008b, page 28) is presented in Table 4, concerning the profile of a special kind of soccer player, and how the several words used to represent it occur in the 50’s, 70’s and 2000’s – and thereafter draw conclusions as to vocabulary trends and convergence/divergence among the varieties at stake (P for Portugal, B for Brazil).

TABLE 4: Absolute and relative frequencies and absolute and relative uniformity U e U' of the AVANÇADO onomasiological profile

Avançado	**P50**		**B50**		**P70**		**B70**		**P00**		**B00**	
atacante	101	8,8	119	36,6	50	13,6	208	73,8	42	9,7	658	96,2
avançado	820	71,6	3	0,9	175	47,4	0	0,0	240	55,4	0	0,0
avante	0	0,0	159	48,9	0	0,0	31	11,0	0	0,0	23	3,4
dianteiro	220	19,2	22	6,8	74	20,1	2	0,7	38	8,8	0	0,0
forward	1	0,1	17	5,2	0	0,0	0	0,0	0	0,0	0	0,0
ponta-de-lança	3	0,3	5	1,5	70	19,0	41	14,5	113	26,1	3	0,4
	U = 16,9		U’ = 0,6		U = 28,8		U’ = 0,8		U = 10,1		U’ = 0,4	

This is a morose process that requires classification of a large number of corpus instances (all occurrences of the forms above). Only after all those decisions have been taken can the measures be computed and compared.

Now, one of the advantages of making the underlying corpora (annotation) available to other researchers is that other people can then inspect the individual classifications, search for the classes and the specific contexts of occurrence, and even provide feedback or corrections if needed. A similar point has been made in Santos & Oksefjell (1999) in what concerns parallel corpora.

This allows for both a wider dissemination of the original research, and a better quality control through communication with one’s peers. Both aims are included in Linguateca’s mission for the computational processing (and study) of the Portuguese language.

[9] From Speelman et al. (2003), *onomasiological variation* concerns “different terms used to refer to the same entity”, while *formal onomasiological variation* requires that no conceptual change is at stake, and therefore does not include cases like hyperonyms or hyponyms which are also frequently used about the same referent in discourse. The authors themselves are aware that this is not easy to distinguish for all corpus instances, though.

It is thus currently possible to ask, in addition to the occurrence or distribution of the forms included in the profiles, for an entire profile, or for the profile distributions themselves. That is, how many cases of the members of the profile CASACO appear by date/decade, or variety.

We have also used the initial profiles compiled in CONDIVport as a seed to compiling larger sets of fashion-related lexical items, thus "colouring" the different corpora also with clothing information.[10]

[3.3] *New capabilities in the AC/DC interface*

Several capabilities newly added to the AC/DC interface deserve mention here:

- Human validation of corpus illustration sentences for semantic relation evaluation (the VARRA service, developed in connection with yet another subproject in Linguateca, PAPEL[11], whose goal was to create a free lexical ontology for Portuguese based on an existing general dictionary);
- Comparison of two search expressions, inspired by the CorpusEye search system (Bick 2004), to compare explicitly two distributions;
- Reuse of a pattern database, inspired by the search system of Davies & Ferreira (2006-) and based on the capabilities of the underlying CWB system (Schulze 1996; Evert 2009).

These will pave the way for yet further developments in the AC/DC cluster, some of which can be mentioned here as natural extensions, namely (i) the possibility to include (tailored) synonym search as an option, following e.g. Christ (1998); and (ii) search by subject matter through concept nets.

Illustration sentences

Although their wealth of real, in context, examples is generally accepted as one of the basic advantages of corpora, as opposed to laboriously crafted ones (by a lexicographer or textbook author), it is not easy to come up automatically with good examples from a corpus, as pointed out by Kilgarriff et al. (2008).

Even harder did we find the task of illustrating, or validating, semantic relations between words in context, as we wished to do for PAPEL, whose relations between words (and not word senses) had been produced automatically and were thus in need of human validation (Gonçalo Oliveira et al. 2009, 2010).

We have thus developed an AC/DC-based service to help us achieve two related purposes: (i) find out the best patterns to validate and/or discover semantic

[10] Whether the use of semantic domains and ontology-based classifications is also useful for variation analysis is something that will have to be ascertained empirically.

[11] See http://www.linguateca.pt/PAPEL/

relations in text, and (ii) develop clearer insights into the semantic fabric of Portuguese, while at the same time improving a public-domain semantic resource. As is common practice in Linguateca, we offer this as a service to the community[12], which means that everyone can use it to develop or evaluate their own resources.

Comparison of two phenomena

Although one could already perform a comparison by doing two (or more) searches in AC/DC on a row and then comparing the results, this capability provides an easier way by aligning the results on two sides of the same screen. Since we have been doing similar things within DISPARA for a long time now, cf. the *quantitative wrapup* function in Santos (2002), it seemed appropriate to offer this also in a monolingual corpus context.

Reuse of a pattern database

Again, this is not new in the sense that in other services offered by Linguateca, namely Águia (Santos 2003), use was made of a set of patterns to query complex treebank structures in the Floresta project, but this feature had never been integrated in the main service interface, which relied mainly in direct e-mail answers to users asking us how to produce complex queries.

Now we have created an option of loading previous queries/commands into the search space, which, although possibly slowing down the corpus system, will also provide higher expressivity. It remains to be seen how much of this will in fact be reused/employed by power users of the AC/DC services.

ACKNOWLEDGEMENTS

Linguateca has throughout the years been jointly funded by the Portuguese Government, the European Union (FEDER and FSE), under contract ref. POSC/339/1.3-/C/NAC, UMIC and FCCN.

I would like to thank the remaining members of the AC/DC team, Paulo Rocha, Luís Costa, Rosário Silva and Cristina Mota for the joint work, the corpus owners for letting us grant access to them on the web, and all users who have requested features or suggested improvements.

Eckhard Bick's long-standing collaboration with his PALAVRAS parser has been the single most important factor for AC/DC's success near its users.

Thanks also to Tony Berber Sardinha and Violeta Quental for relevant information concerning the history of Brazilian corpora, to Fernanda Bacelar do Nascimento for relevant references, to Augusto Soares da Silva for the introduction to quantitative lexicology methodology, to Cristina Mota for pertinent comments on a draft version and, last but not least, to the VARRA team (Cláudia Freitas, Hugo

[12] See http://www.linguateca.pt/acesso/varra.php

Gonçalo Oliveira and Violeta Quental).

references

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002. Floresta Sintá(c)tica: a treebank for Portuguese. In Manuel Gonzalez Rodrigues & Carmen Paz Suarez Araujo (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, 1698–1703. Paris: ELRA.

Aluísio, Sandra, Gisele Montilha Pinheiro, Aline M.P. Manfrin, Leandro H.M. de Oliveira, Luiz C. Genoves Jr & Stella E. O. Tagnin. 2004. The Lácio-Web: Corpora and tools to advance Brazilian Portuguese language investigations and computational linguistic tools. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1779–1782.

Bacelar do Nascimento, Maria Fernanda. 2000. O corpus de referência do português contemporâneo e os projectos de investigação do Centro de Linguística da Universidade de Lisboa sobre variedades do português falado e escrito. In E. Gärtner, C. Hundt & A. Schönberger (eds.), *Estudos de gramática portuguesa (I)*, 185–200. Centro do Livro e do Disco de Língua Portuguesa. Biblioteca Luso-Brasileira.

Bacelar do Nascimento, Maria Fernanda. 2001. Les études portugaises sur la langue parlée. In M. H. A. Carreira (ed.), *Travaux et documents, les langues romanes en dialogue(s)*, vol. 11, 209–221. Université Paris 8.

Bacelar do Nascimento, Maria Fernanda, Antónia Estrela, Amália Mendes, Luisa Pereira & Rita Veloso. 2008a. African Varieties of Portuguese: Corpus Constitution and Lexical Analysis. In *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics - LULCL II 2008*.

Bacelar do Nascimento, Maria Fernanda, M. L. G. Marques & Maria Luísa Segura da Cruz. 1984. *Português Fundamental: Vocabulário e Gramática*. Centro de Linguística da Universidade de Lisboa.

Bacelar do Nascimento, Maria Fernanda, M. L. G. Marques & Maria Luísa Segura da Cruz. 1987. *Português Fundamental: Métodos e Documentos*, vol. 2. Centro de Linguística da Universidade de Lisboa.

Bacelar do Nascimento, Maria Fernanda, Luisa Pereira, Antonia Estrela, José Bettencourt Gonçalves & Sancho Oliveira. 2008b. Aspectos de unidade e diversidade do português: as variedades africanas face à variedade europeia. *Veredas* 9. 35–69.

Bacelar do Nascimento, Maria Fernanda, Maria Celeste Rodrigues & José Bettencourt Gonçalves (eds.). 1996. *Actas do XI Encontro Nacional da associação portuguesa de linguística (Lisboa, 2-4 de Outubro de 1995), vol I: Corpora*. Lisboa, Portugal: APL/Colibri.

Barreiro, Anabela, Luzia Helena Wittmann & Maria de Jesus Pereira. 1996. Lexical differences between European and Brazilian Portugueses. *INESC Journal of Research and Development* 5(2). 75–101.

Berber Sardinha, A. P. 1999. Beginning Portuguese corpus linguistics: exploring a corpus to teach Portuguese as a foreign language. *DELTA* 15(2). 289–299.

Berber Sardinha, Tony. 2007. History and compilation of a large register-diversified corpus of Portuguese at CEPRIL. *The Especialist* 28. 211–226.

Berber Sardinha, Tony & Gladis Maria de Barcellos Almeida. 2008. A Lingüística de Corpus no Brasil. In Stella E. O. Tagnin & Oto Araújo Vale (eds.), *Avanços da Lingüística de Corpus no Brasil*, 17–40. Editora Humanitas/FFLCH/USP.

Bick, Eckhard. 1997. Internet Based Grammar Teaching. In Ellen Christoffersen & Bradley Music (eds.), *Proceedings of Datalingvistisk Forenings Årsmøde 1997 i Kolding*, 86–106. Handelshøjskole Syd, Institut for Erhvervssprog og Sproglig Informatik.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Ph.D. thesis, Aarhus University, Aarhus, Denmark.

Bick, Eckhard. 2004. Looking at the Floresta Sintá(c)tica with a CorpusEye: A user-friendly cross-language search interface. URL http://www.linguateca.pt/documentos/floresta-corpuseye_en.pdf.

Brandão, Silvia Figueiredo & Maria Antónia Mota (eds.). 2003. *Análise contrastiva de variedades do português: primeiros estudos*. In-Fólio.

Callou, Dinah. 1999. O projecto NURC no Brasil: da década de 70 à década de 90. *Linguística* 11. 231–250.

Castilho, Ataliba Teixeira de, Gisele Machline de Oliveira e Silva & Dante Lucchesi. 1995. Informatização de acervos da língua portuguesa. *Boletim da Abralin* 17. 143–151.

Christ, Oliver. 1998. Linking WordNet to a Corpus Query System. In John Nerbonne (ed.), *Linguistic databases*, CSLI lecture notes, 189–202. CSLI Publications, CSLI Stanford.

Davies, Mark. 2008. New Directions in Spanish and Portuguese Corpus Linguistics. *Studies in Hispanic and Lusophone Linguistics* 1. 149–186.

Davies, Mark & Michael Ferreira. 2006-. Corpus do português. URL http://www.corpusdoportugues.org. 45 milhões de palavras, sécs. XIV-XX.

Evert, Stefan. 2009. The CQP Query Language Tutorial. URL http://cwb.sourceforge.net/temp/CQPTutorial.pdf.

Frankenberg-Garcia, Ana & Diana Santos. 2003. Introducing COMPARA, the Portuguese-English parallel translation corpus. In Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in Translation Education*, 71–87. Manchester: St. Jerome Publishing.

Freitas, Cláudia, Paulo Rocha & Eckhard Bick. 2008. Floresta Sintá(c)tica: Bigger, Thicker and Easier. In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)*, 216–219. Berlin/Heidelberg: Springer Verlag.

Galves, Charlotte. 2009. Padrões rítmicos, domínios prosódicos e modelagem probabilística em corpora do português. URL http://www.tycho.iel.unicamp.br/tycho/prdpmp/projetofinal.pdf.

Geeraerts, Dirk & Stefan Grondelaers. 1999. Purism and fashion. French influence on Belgian and Netherlandic Dutch. *Belgian Journal of Linguistics* 13. 53–68.

Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de nederlandse woordenschat*. Amsterdam: Meertens Instituut.

Gonçalo Oliveira, Hugo, Diana Santos & Paulo Gomes. 2009. Relations extracted from a Portuguese dictionary: results and first evaluation. In Luís Seabra Lopes, Nuno Lau, Pedro Mariano & Luís M. Rocha (eds.), *New Trends in Artificial Intelligence, Local Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA 2009)*, 541–552.

Gonçalo Oliveira, Hugo, Diana Santos & Paulo Gomes. 2010. Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação. *Linguamática* 2(1). 77–93.

Gougenheim, G., R. Michéa, P. Rivenc & A. Sauvageot. 1964. *L'élaboration du français fondamental*. Paris: Didier.

Inácio, Susana & Diana Santos. 2006. Syntactical Annotation of COMPARA: Workflow and First Results. In Renata Vieira, Paulo Quaresma, Maria da Graça

Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006*, 256–259. Berlin/Heidelberg: Springer Verlag.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell & Pavel Rychlý. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX 2008*. Barcelona.

Maciel, Carlos. 1997. Textes et textes juridiques dans la Base de Données Textuelles PORTEXT. In *Secondes Journées Internationales de Terminologie (Actes du colloque, Le Havre, 14-15 octobre 1994)*. Le Havre.

Menon, Odete Pereira da Silva, Edson Domingos Fagundes & Loremi Loregian-Penkal. 2009. The VARSUL database. *Linguistik online* 38(2). 13–21.

Oksefjell, Signe. 1999. A Description of the English-Norwegian Parallel Corpus: Compilation and Further Developments. *International Journal of Corpus Linguistics* 4(2). 197–216.

Oksefjell, Signe & Diana Santos. 1998. Breve panorâmica dos recursos de português mencionados na Web. In Vera Lúcia Strube de Lima (ed.), *III Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR'98)*, 38–47.

Peres, João Andrade & Mary Aizawa Kato. 2004. Studies in the Comparative Semantics of European and Brazilian Portuguese. Special issue of *Journal of Portuguese Linguistics*, 3 (1).

Santos, Diana. 1999a. Computational processing of Portuguese: working memo. URL http://www.linguateca.pt/branco/white_paper.html.

Santos, Diana. 1999b. Disponibilização de corpora de texto através da WWW. In Palmira Marrafa & Maria Antónia Mota (eds.), *Linguística Computacional: Investigação Fundamental e Aplicações. I Workshop sobre Linguística Computacional da APL, FLUL, Maio de 1998*, 323–335. Lisboa: Colibri / APL.

Santos, Diana. 2002. DISPARA, a system for distributing parallel corpora on the Web. In Nuno Mamede & Elisabete Ranchhod (eds.), *Advances in Natural Language Processing: Third International Conference, Proceedings (PorTAL 2002)*, 209–218. Berlin/Heidelberg: Springer-Verlag.

Santos, Diana. 2003. Timber! Issues in treebank building and use. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language: 6th International Workshop,* *PROPOR 2003. Faro, Portugal, June 2003*, 151–158. Berlin/Heidelberg: Springer Verlag.

Santos, Diana. 2009. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamatica* 1(1). 25–59.

Santos, Diana & Cristina Mota. 2010. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, 1437–1444. European Language Resources Association.

Santos, Diana & Signe Oksefjell. 1999. Using a parallel corpus to validate independent claims. *Languages in Contrast* 2(1). 117–132.

Santos, Diana & Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 442–449.

Schulze, Maximilian Bruno. 1996. *MP User's Manual.* Institutt für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Silva, Rosário & Diana Santos. 2009. Arco-íris: notas sobre a anotação do campo semântico da cor em português. URL http://www.linguateca.pt/acesso/ArcoIris.pdf. First version: 25 June 2009.

Soares da Silva, Augusto. 2008a. Integrando a variação social e métodos quantitativos na investigação sobre linguagem e cognição: para uma sociolinguística cognitiva do português europeu e brasileiro. *Revista de Estudos Linguísticos* 16(1). 49–81.

Soares da Silva, Augusto. 2008b. O corpus CONDIV e o estudo da convergência e divergência entre variedades do português. In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*, 25–28. Linguateca.

Soares da Silva, Augusto. 2010. Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In Dirk Geeraerts, Gitte Kristiansen & Yves Peirsman (eds.), *Advances in cognitive sociolinguistics*, 41–83. Mouton de Gruyter.

Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts. 2003. Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37. 317–337.

Tagnin, Stella E. O., Elisa Duarte Teixeira & Diana Santos. 2009. CorTrad: a multiversion translation corpus for the Portuguese-English pair. *Arena Romanistica* 4. 314–323. [The 28th International Conference on lexis and grammar, Bergen, Norway, 30 September - 3 October 2009].

Thomson, H., S. Armstrong-Warwick & D. McKelvie. 1994. Data in your language: The ECI Multilingual Corpus 1. In *Proceedings of the International Workshop on Shareable Natural Language Resources (Nara, Japan, 10-11 August 1994)*. Institute of Science and Technology.

Varejão, Filomena de Oliveira Azevedo. 2009. O português do Brasil: Revisitando a História. *Cadernos de Letras da UFF – Dossiê: Difusão da língua portuguesa* 39. 119–137.

Wittmann, Luzia, Tânia Pêgo & Diana Santos. 1995. Português do Brasil e de Portugal: alguns contrastes. In *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística*, 465–487. Lisboa: APL/Colibri.

AUTHOR CONTACT INFORMATION

Diana Santos
Department of Literature, Area Studies and European Languages
Faculty of Humanities, University of Oslo
P.O.Box 1003 Blindern
N-0315 Oslo
Norway
d.s.m.santos@ilos.uio.no

J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 129–145. (ISSN 1890-9639)
http://www.journals.uio.no/osla

THE HISPACAT COMPARATIVE DATABASE OF SYNTACTIC CONSTRUCTIONS AND ITS APPLICATIONS TO SYNTACTIC VARIATION RESEARCH

XAVIER VILLALBA
Barcelona

ABSTRACT

The HISPACAT database of syntactic constructions in Catalan and Spanish is a dynamic comparative grammar of two closely related languages, which, from a theoretical point of view, offers us a alternative viewpoint to advance in our understanding of the "atoms" of linguistic microvariation, and to offer a snapshot of microparametric (in)variance, which will help us to predict less stable parts of the grammar, and hence more sensitive to syntactic change or interference phenomena. Moreover, this tool, which is conceived as a major empirical source for testing syntactic microvariation, may also prove helpful for researchers in bilingualism and language contact studies, and for teachers and students of Catalan or Spanish as L2.

[1] INTRODUCTION

HISPACAT, the database of syntactic constructions in Catalan and Spanish, integrates in a bigger funded research project of the Centre de Lingüística Teòrica of the Universitat Autònoma de Barcelona, which involves a team of 16 senior and 10 junior researchers, and one head technician, Daniel Jiménez, headed by professors Carme Picallo and Josep M. Brucart.[1]

HISPACAT is conceived as a major empirical source for detecting syntactic microvariation, and testing hypotheses regarding bilingualism, and interference, but which may prove helpful for L2 learning as well, for it is designed as a dynamic comparative grammar of two closely related languages capable of offering

[1] I am really thankful to the organizers and the audience of the Workshop on Research Infrastructure for Linguistic Variation for the lively atmosphere of knowledge-sharing, and illuminating discussion. Thanks are also due to the comments of three anonymous reviewers, which have improved the paper significantly. The research included in this paper has been supported by the funded research projects BFF2003-08364-C02 (Spanish MCyT and EU FEDER) and HUM2006-13295-C02/FILO (Spanish MEC and EU FEDER) awarded to the Centre de Lingüística Teòrica of the Universitat Autònoma de Barcelona.

a snapshot of basic (in)variance patterns. HISPACAT aims to help researchers investigating what factors of the computational system and what morphosyntactic features of lexical expressions underlie the grammatical symmetries and asymmetries between Spanish and Catalan. Moreover, as the unchanging principles governing language can only be fully explained in connection to a proper understanding of the nature of linguistic variation, HISPACAT is intended to be a modest but valuable tool for building a general theory of language.

In the first part of the paper we will discuss the goals and theoretical foundations of the project, stressing the theoretical and methodological differences existing between our enterprise and current textual databases. In the second part of the paper, a general presentation will be offered of the main architecture of the database, namely the internal structure of the files, and its conceptual ontology. It will be shown that the conceptual-based design of HISPACAT, besides providing an exhaustive description of the basic features of each construction, will prove instrumental for allowing a flexible system of information queries.

Finally, the last part of the paper will be devoted to analyze several cases related to microvariation, and interference phenomena, where HISPACAT has proved to be a reliable source for establishing robust linguistic generalizations.

[2] GOALS AND THEORETICAL FOUNDATIONS OF HISPACAT

HISPACAT was conceived as a tool within a theory-driven project aimed at the identification of the "atoms" underlying syntactic microvariation in Catalan and Spanish and at the prediction of domains most vulnerable to syntactic interference. Moreover, HISPACAT is an applied project aimed at building an empirical playground for researchers in bilingualism and L2 learning and a dynamic comparative grammar for L2 teachers/learners.

[2.1] *Theoretical foundations*

From a theoretical point of view, the work by Richard S. Kayne – see Kayne (1996, 2000, 2005); see also the pioneering work by Borer (1984), and Rizzi (1982) – has emphasized that we must go beyond the concept of (macro)parameter, coined to give account of large structural differences between language groups – like the null subject (Jaeggli & Safir (1986)) or the polysynthetic parameter (Baker (1996)) – and focus attention on variation at a smaller scale, namely minor inter or crosslinguistic nuances not affecting the overall typological profile of languages, but nonetheless generating significant differences in the behavior of certain linguistic units. This microparametric framework is thus the most suitable for the comparative work on which HISPACAT is grounded.

Just to emphasize the theoretical and empirical linguistic relevance of the HISPACAT project, we will briefly consider the constructions involving directive modality in HISPACAT. The common core includes the imperative mood for affirmative orders (1) and the subjunctive mood for prohibitions (2), the translated use of present (3) and future tenses (4), and the exhortative subjunctive mood introduced by the complementizer (5):

(1) a. Tanca la porta.
close the door
'Close the door!'
(Catalan: (Payrató 2002, 3.4.4))

b. ¡Cállate, estúpido!
shut.up stupid
'Shut up, you idiot!'
(Spanish: (Garrido 1999, 60.2.1.5))

(2) a. ¡No diguis aquestes coses!
not say.SUBJUNCTIVE.2SG these things
'Don't say that kind of things!'
(Catalan: (Espinal 2002, 24.2.2))

b. No se lo des.
not him it give.2SG
'Don't give it to him!'
(Spanish: (Garrido 1999, 60.2.1.3)

(3) a. ¡Ho fas i santes pasqües!
it do.2SG and holy.PL Easters
'Do it, period!'
(Catalan: (Payrató 2002, 3.4.4))

b. De noche sales conmigo o no sales.
of night go.out.2SG with.me or not go.out.2SG
'At night, you go out with me or you don't go out.'
(Spanish: (Fernández-Ramírez 1951, V.46))

(4) a. No mataràs.
not kill.FUT.2SG
'You shall not murder.'
(Catalan: (Saldanya 2002, 22.5.7.3))

b. ¿y me vas a dejar sola? ¡oh! No harás tal cosa.
and me go.you to leave alone o not do.future.2SG such thing
'and you are gonna leave me alone? Why, you won't!'
(Spanish: (Fernández-Ramírez 1951, V.46))

(5) a. ¡Que ho hagin dibuixat abans de plegar!
that it have.SUBJUNCTIVE.3pl draw before of leave
'They shall draw it before leaving!'
(Catalan: (Payrató 2002, 3.4.4))

b. ¡Que venga Juan!
that come.SUBJUNCTIVE.3SG Juan
'Juan shall come!'
(Spanish: (Ridruejo 1999, 49.1.3))

The contrast between Catalan and Spanish arises concerning the use of infinitives for conveying directives. Even though both languages share the prepositional construction in (6), they sharply contrast with respect to infinitival (7) and retrospective imperatives (8):

(6) a. Va, nois, a treballar. (Catalan)
come.3PL boys to work
'Come on, boys, to the job!'
b. Venga, chicos, a trabajar.
come.2SG boys to work
'Come on, boys, to the job!'
(Spanish: (Hernanz 1999, 36.4.2.3))

(7) a. *Nens, ¡fer-me cas! (Catalan)
kids make.me case
'Kids, pay me attention!'
b. Niños, ¡hacerme caso!
kids make.me case
'Kids, pay me attention!'
(Spanish: (Hernanz 1999, 36.4.2.3))

(8) a. ??¡Haver-ho dit abans!
have-it said before
'You should have said it before!'
(Catalan: (Payrató 2002, 3.4.4.2))
b. Siento mucho llegar tarde. Haber salido antes de casa.
feel much arrive late have left before of house
'I am very sorry for being late. You should have leave home earlier.'
(Spanish: (Garrido 1999, 60.2.1.6))

Interestingly, the HISPACAT database shows us that Catalan resorts to subjunctive in these cases:

(9) a. ¡Ho haguessis dit abans!
it have.PAST.SUBJUNCTIVE.2SG said before
'You should have said it before!'
(Catalan: (Payrató 2002, 3.4.4.2))

b. *¡Lo hubieses dicho antes! (Spanish)
it have.PAST.SUBJUNCTIVE.2SG said before
'You should have said it before!'

Furthermore, Catalan makes a distinctive use of the subjunctive in related constructions both affirmative (10) and (11) negative, which suggests a very specific area of microvariation:

(10) a. La tornin.
her return.2PL
'Bring it back.'
(Catalan: (Payrató 2002, 3.4.4.1))
b. *La devuelvan. (Spanish)
her return.2PL
'Bring it back.'

(11) a. No hi baixéssiu pas, sentiu?
not here come.down not hear.2PL
'Don't move downwards, ok?'
(Catalan: (Quer 2002, 22.6.3.2))
b. *¡No bajaseis, oís? (Spanish)
not come.down hear.2PL
'Don't move downwards, ok?'

The resultant picture suggests as well hypotheses concerning interference phenomena in this area of grammar. For instance, since Catalan and Spanish share the correlation between affirmative directives with imperative mood (1) and between negative directives with subjunctive mood (2), we don't expect the extension of the Spanish directive infinitive to Catalan grammar.

We must remark that the relevant evidence was there, but buried into, and scattered through, two separate monolingual grammars. HISPACAT proves, thus, to be particularly useful at bringing all the pieces together in the form of a relational database, facilitating deeper empirical generalizations and far-reaching hypotheses.

[2.2] *Linguistically-oriented design*

The strongly comparative and linguistically-oriented nature of this project extends to its applied design as well, so that HISPACAT comes to fill a gap in the field of Catalan and Spanish linguistic databases. Even though we count with academic textual databases like the Catalan *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC) and the Spanish *Corpus de Referencia del Español Actual* (CREA) or *Archivo Gramatical de la Lengua Española* (AGLE), concerning syntactic microvariation, they suffer from three major drawbacks:

- they are monolingual,
- they focus on the data rather than on grammatical concepts (with the noteworthy exception of Spanish AGLE), and
- they do not include negative data.

The HISPACAT database, in contrast,

- focuses on comparative data,
- is a relational database focused on the grammatical features underlying linguistic phenomena, and
- includes negative data that set the limits of grammaticality of constructions.

Regarding the first aspect, HISPACAT is designed and built in a totally contrastive way: the presentation of data aims at showing the contact points *and* the syntactic asymmetries between Catalan and Spanish, an enterprise which has not been attempted to date. We are, thus, interested in data like the following (see also the discussion in 2.1 above):

(12) Tinc por que no vinguin. (Catalan)
have fear that not come.SUBJUNCTIVE.3PL
a. negative: 'I am afraid that they won't come.'
b. expletive: 'I am afraid that they should come.'

(13) Tengo miedo de que no vengan. (Spanish)
have fear of that not come.SUBJUNCTIVE.3PL
a. negative: 'I am afraid that they won't come.'
b. expletive: *'I am afraid that they should come.'

Whereas Catalan makes an ambiguous use of negation in this particular context, Spanish can only obtain the negative reading. Our main concern is determine why this is so and try to connect this fact with other apparently unrelated ones to help us understand the key factors underlying variation in this area of grammar. Therefore, we can say that, despite its database format, HISPACAT is a dynamic comparative grammar of Catalan and Spanish, which attempts to connect the valuable independent findings arrived at by the two major grammatical works of Spanish and Catalan: Bosque & Demonte (1999), and Solà et al. (2002).

Regarding the second aspect, HISPACAT is conceived as a relational database based on linguistic concepts, rather than on particular occurrences from a textual corpus, for we are interested not in the linguistic expressions themselves but on the concepts underlying them. Obviously, this line of work presupposes the theoretical hypotheses that (micro)syntactic variation is the result of the setting of (micro)parameters – see Kayne (1996, 2000, 2005) – and that syntactic constructions are not primitive but the result of a sum of properties – Chomsky (1981); cf. the basic assumptions of Construction Grammar, as developed by Goldberg (1995).[2]

Finally, also in contrast to current textual corpora, HISPACAT allows the inclusion of negative data helping the reader to appreciate the grammaticality limits of constructions. For instance, in the file corresponding to the file *expletive negation in the context of comparative markers*, we include (and comment on) data like the following:

(14) a. Juan era antes más simpático que (no) ahora. (Spanish)
Juan was before more friendly that not now
'Juan was more friendly then than now.'
b. Juan era antes más simpático de lo que (*no) es ahora. (Spanish)
Juan was before more friendly of it that not is now
'Juan was more friendly then than now.'

Anyone familiar with L2 learning will appreciate the advantages of including such information, unavailable in textual corpora: the provision of information about what is possible, but also about what is not (a typical shortcoming of school grammars), helps the teacher to set the properties and extent of any construction in a more precise way, while helping learners to avoid incorrect generalizations.

[2.3] *Applied goals*

From an applied point of view, the HISPACAT database has three main goals:

(i) providing an empirical database for research on bilingualism and L2 learning,

(ii) providing a comparative grammar for teachers and students of Catalan or Spanish as L2,

(iii) providing a catalog of commented examples for teachers and students of Catalan or Spanish as L2.

There is no doubt that an increasing number of researchers are approaching the phenomena of language contact, bilingualism, and L2 learning from a gram-

[2] Certainly, as one reviewer acutely remarks, the limits between constructions can hardly be stated precisely. We agree with this observation, which reflects the empirical nature of the HISPACAT database: the actual list of constructions is not a closed set, but is subject to continuous revision in the light of new evidence.

matical point of view. For these researchers, HISPACAT will prove instrumental for contributing to refine the descriptions and hypotheses of scholars in such areas of applied linguistics (goal 1). Moreover, HISPACAT can also be useful to teachers and students of Catalan or Spanish as a L2, since it offers a dynamic user-oriented comparative grammar of these two languages (goal 2). Finally, the inclusion of a big number of analyzed examples of both grammatical and ungrammatical constructions represents an added value for teachers and students of Catalan or Spanish as L2, for it increases the consistency of grammatical description and facilitates the development of specific teaching materials (goal 3).

[3] ARCHITECTURE OF HISPACAT

HISPACAT is a build as an ORACLE concept-oriented relational database with two key design features:

- a system of files built on a comparative basis, and including both descriptive and analytical fields, and
- an ontology of linguistic concepts, aimed at offering both a detailed description of constructions, and a robust data-retrieval system.

[3.1] *The files*

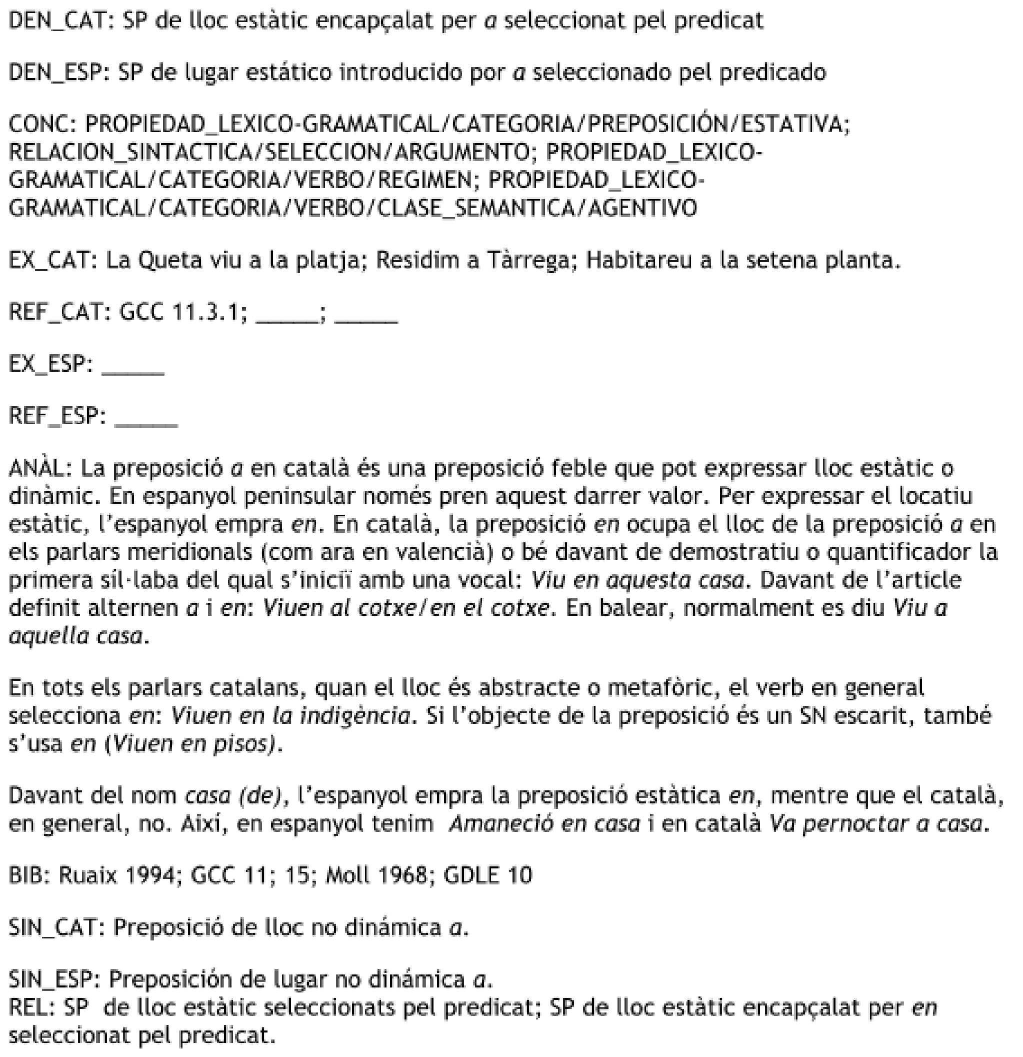
HISPACAT is fed through a file system including a series of fields designed to emphasize the comparative and conceptual-based nature of the database. In Fig. 1 on the facing page you have a (simplified) example of the file corresponding to *selected static locative PP headed by a 'to'*:[3]

[3.2] *The ontology*

In order to help researchers to discover the basic properties underlying each construction, an ontology was designed of 176 linguistic concepts, classified into relationships and properties, and further subdivided into lexical-grammatical and semantic-pragmatic properties and semantic and syntactic relations, as shown in Fig. 2 on page 138.

Moreover, every concept is associated with a code based on their position in the ontology, which allows us to show direct relationship networks and natural classes of concepts. Henceforth, the files of the database are designed as comparative concept-based descriptions of constructions, as in the following example (the lack of Spanish examples indicates that the construction is available in Catalan only).

[3] The codes are the following: DEN-CAT=Catalan term, DEN-ESP=Spanish term, CONC=linguistic concepts, EX-CAT=Catalan examples, REF-CAT=reference of Catalan examples, EX-ESP=Spanish examples, REF-ESP=reference of Spanish examples, ANAL=analysis, BIB=bibliography, SIN-CAT=synonyms of Catalan term, SIN-ESP=synonyms of Spanish term, and REL=related constructions.



DEN_CAT: SP de lloc estàtic encapçalat per *a* seleccionat pel predicat

DEN_ESP: SP de lugar estático introducido por *a* seleccionado pel predicado

CONC: PROPIEDAD_LEXICO-GRAMATICAL/CATEGORIA/PREPOSICIÓN/ESTATIVA; RELACION_SINTACTICA/SELECCION/ARGUMENTO; PROPIEDAD_LEXICO-GRAMATICAL/CATEGORIA/VERBO/REGIMEN; PROPIEDAD_LEXICO-GRAMATICAL/CATEGORIA/VERBO/CLASE_SEMANTICA/AGENTIVO

EX_CAT: La Queta viu a la platja; Residim a Tàrrega; Habitareu a la setena planta.

REF_CAT: GCC 11.3.1; _____; _____

EX_ESP: _____

REF_ESP: _____

ANÀL: La preposició *a* en català és una preposició feble que pot expressar lloc estàtic o dinàmic. En espanyol peninsular només pren aquest darrer valor. Per expressar el locatiu estàtic, l'espanyol empra *en*. En català, la preposició *en* ocupa el lloc de la preposició *a* en els parlars meridionals (com ara en valencià) o bé davant de demostratiu o quantificador la primera síl·laba del qual s'iniciï amb una vocal: *Viu en aquesta casa*. Davant de l'article definit alternen *a* i *en*: *Viuen al cotxe/en el cotxe*. En balear, normalment es diu *Viu a aquella casa*.

En tots els parlars catalans, quan el lloc és abstracte o metafòric, el verb en general selecciona *en*: *Viuen en la indigència*. Si l'objecte de la preposició és un SN escarit, també s'usa *en* (*Viuen en pisos)*.

Davant del nom *casa (de)*, l'espanyol empra la preposició estàtica *en*, mentre que el català, en general, no. Així, en espanyol tenim *Amaneció en casa* i en català *Va pernoctar a casa*.

BIB: Ruaix 1994; GCC 11; 15; Moll 1968; GDLE 10

SIN_CAT: Preposició de lloc no dinámica *a*.

SIN_ESP: Preposición de lugar no dinámica *a*.
REL: SP de lloc estàtic seleccionats pel predicat; SP de lloc estàtic encapçalat per *en* seleccionat pel predicat.

FIGURE 1: Example of HISPACAT file.

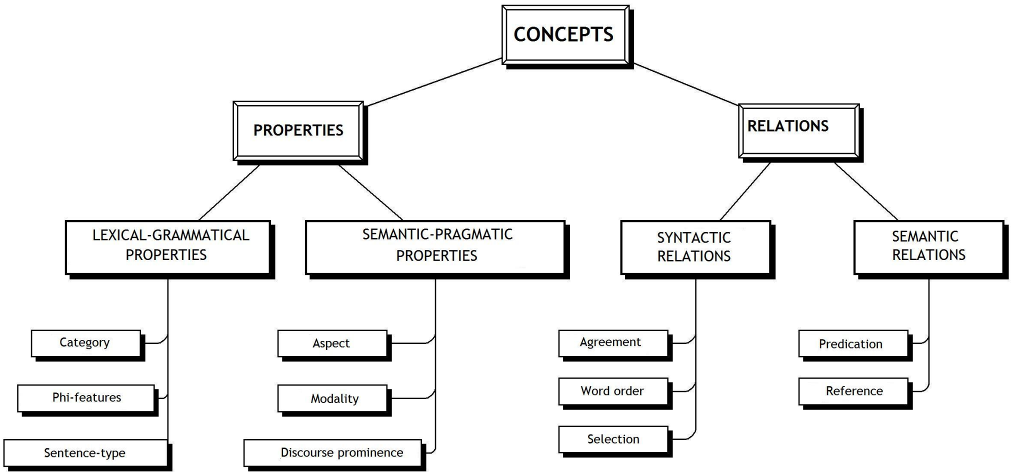


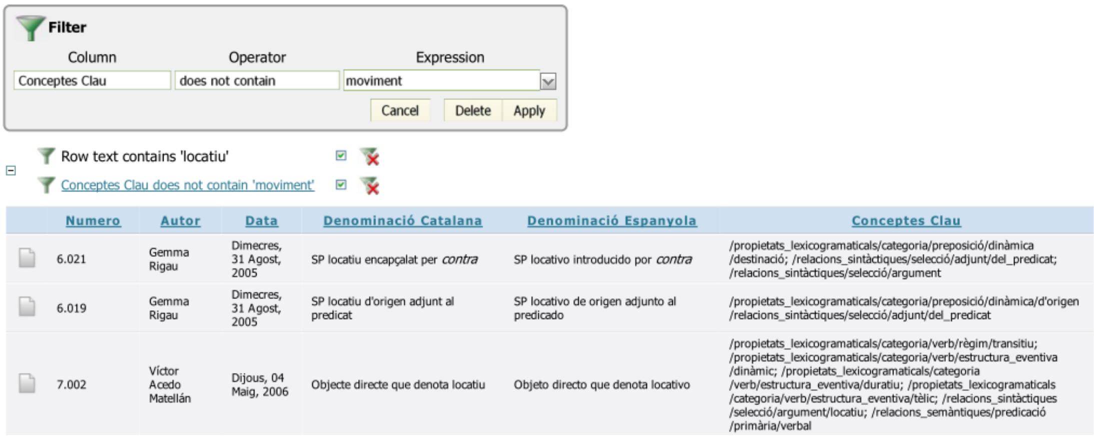
FIGURE 2: Outline of the HISPACAT ontology

(15) Selected Static locative PP headed by *a* 'to'

- EX-CAT: Residim a Tàrrega. ('We live in Tàrrega.'); Habitareu a la setena planta. ('You will dwell in the seventh floor.')
- EX-ESP:
- CONC:
  (i) lexical-grammatical-property / category / preposition / static
  (ii) lexical-grammatical-property / category / verb / selection / inergative
  (iii) lexical-grammatical-property / category / verb / eventive-structure / stative
  (iv) syntactic-relation / selection / argument
- ANAL: Catalan preposition *a* 'to' is a weak preposition that may express a static location or movement. In Peninsular Spanish only the last value is possible. To express the static locative meaning, Spanish must resort to *en* 'in'. When preceding a demonstrative or quantifier beginning with an initial vowel, Catalan changes to *en* 'in': *Viu en aquesta casa* 'Lives at this house.' In front of the definite article, *a* 'to' and *en* 'in' alternate: *Viuen al cotxe/en el cotxe* 'They live at the car.' In Catalan, when the place is metaphorical, the verb selects *en* 'in': *Viuen en la indigència* 'They live amidst poverty'. If the argument of the locative preposition is a bare noun, the preposition *en* 'in' is chosen (Viuen en pisos 'They live in flats.'). When preceding the name *casa* 'home', Spanish selects the static preposition *en* 'in', whereas Catalan

resorts to *a* ‘to’: Sp. *Amaneció en casa* ‘(S)he saw the new day at home’ and Cat. *Va pernoctar a casa* ‘(S)he stay the night at home’.

Importantly, besides its theoretical grounds, the CONC field provides HISPACAT with a strong tool for carrying out complex conceptual searches. For example, HISPACAT allows the user to get the list of the constructions sharing, for example, the concept *locative* while not including the concept *movement* (Fig. 3):



Filter

Column	Operator	Expression
Conceptes Clau	does not contain	moviment

Row text contains 'locatiu' ☑
Conceptes Clau does not contain 'moviment' ☑

Numero	Autor	Data	Denominació Catalana	Denominació Espanyola	Conceptes Clau
6.021	Gemma Rigau	Dimecres, 31 Agost, 2005	SP locatiu encapçalat per *contra*	SP locativo introducido por *contra*	/propietats_lexicogramaticals/categoria/preposició/dinàmica/destinació; /relacions_sintàctiques/selecció/adjunt/del_predicat; /relacions_sintàctiques/selecció/argument
6.019	Gemma Rigau	Dimecres, 31 Agost, 2005	SP locatiu d'origen adjunt al predicat	SP locativo de origen adjunto al predicado	/propietats_lexicogramaticals/categoria/preposició/dinàmica/d'origen /relacions_sintàctiques/selecció/adjunt/del_predicat
7.002	Víctor Acedo Matellán	Dijous, 04 Maig, 2006	Objecte directe que denota locatiu	Objeto directo que denota locativo	/propietats_lexicogramaticals/categoria/verb/règim/transitiu; /propietats_lexicogramaticals/categoria/verb/estructura_eventiva/dinàmic; /propietats_lexicogramaticals/categoria/verb/estructura_eventiva/duratiu; /propietats_lexicogramaticals/categoria/verb/estructura_eventiva/tèlic; /relacions_sintàctiques/selecció/argument/locatiu; /relacions_semàntiques/predicació/primària/verbal

FIGURE 3: Boolean search.

Moreover, beyond standard boolean searches, the conceptual-based design of the database allows us to inspect the files through a conceptual tree based on the ontology. Thus, we can move from the top of the tree to a particular node and get the files including the concept corresponding to this node (Fig. 4 on the next page).

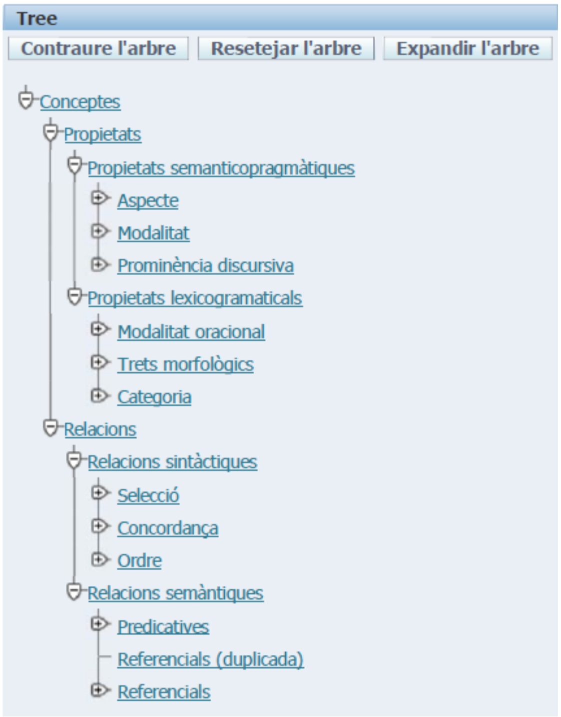
It must be emphasized that the rich system of conceptual searches is complemented with the possibility of textual searches, incorporating thus the advantages of textual corpora (Fig. 5 on page 141).

[4] HISPACAT AT WORK

After this quick glance at the basic architectural features of the HISPACAT database, now we will briefly discuss some immediate applications both at the theoretical and applied level.

[4.1] *Syntactic microvariation: theory and practice*

HISPACAT was originally designed as a tool for the study of syntactic microvariation, which is the main interest of the general research project. Its development has crucially raised many insights in this area, as we have discussed in 2.1. Now consider another case, that of the construction SELECTED STATIC LOCATIVE PP



- Conceptes
  - Propietats
    - Propietats semanticopragmàtiques
      - Aspecte
      - Modalitat
      - Prominència discursiva
    - Propietats lexicogramaticals
      - Modalitat oracional
      - Trets morfològics
      - Categoria
  - Relacions
    - Relacions sintàctiques
      - Selecció
      - Concordança
      - Ordre
    - Relacions semàntiques
      - Predicatives
      - Referencials (duplicada)
      - Referencials

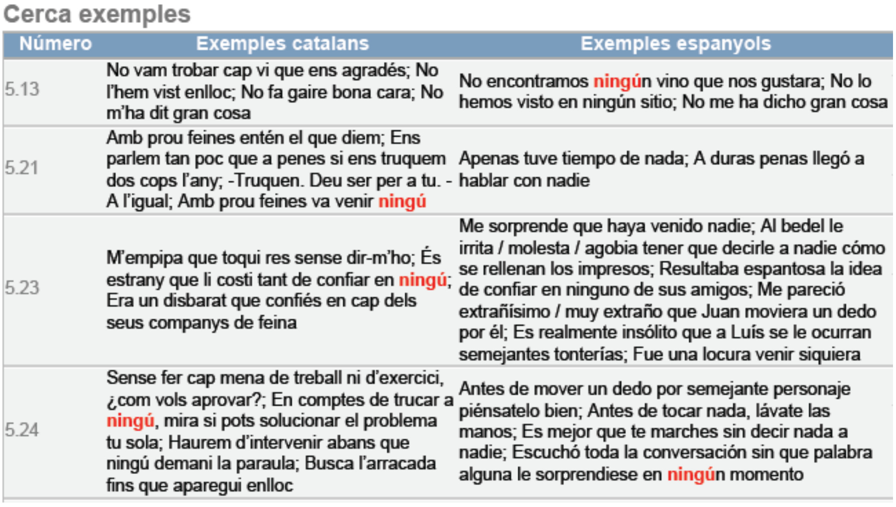
FIGURE 4: The conceptual tree.

HEADED BY A 'TO', which has been presented before in (15):[4]

(16) Selected Static locative PP headed by *a* 'to'

- EX-CAT: Residim a Tàrrega. ('We live in Tàrrega.'); Habitareu a la setena planta. ('You will dwell in the seventh floor.')
- EX-ESP:
- ANAL: Catalan preposition *a* 'to' is a weak preposition that may express a static location or movement. In Peninsular Spanish only the last value is possible. To express the static locative meaning, Spanish must resort to *en* 'in'. When preceding a demonstrative or quantifier beginning with an initial vowel, Catalan changes to *en* 'in': Viu en aquesta casa 'Lives at this house.' In front of the definite article, *a* 'to' and *en* 'in' alternate: Viuen al cotxe/en el cotxe 'They live at the car.' In Catalan, when the place is metaphorical, the verb selects *en* 'in': Viuen en la indigència 'They live

[4] This brief example is not intended to exhaust the issue. As one reviewer points out, a more accurate description would be that "in Spanish 'a' can be locative, but only refers to the contact with a line (a perimeter or otherwise), while in Catalan the same preposition can refer to this type of contact or to the inclusion of an object inside the area of a bidimensional object". For a detailed discussion, see Fábregas (2007).



Cerca exemples

Número	Exemples catalans	Exemples espanyols
5.13	No vam trobar cap vi que ens agradés; No l'hem vist enlloc; No fa gaire bona cara; No m'ha dit gran cosa	No encontramos ningún vino que nos gustara; No lo hemos visto en ningún sitio; No me ha dicho gran cosa
5.21	Amb prou feines entén el que diem; Ens parlem tan poc que a penes si ens truquem dos cops l'any; -Truquen. Deu ser per a tu. -A l'igual; Amb prou feines va venir ningú	Apenas tuve tiempo de nada; A duras penas llegó a hablar con nadie
5.23	M'empipa que toqui res sense dir-m'ho; És estrany que li costi tant de confiar en ningú; Era un disbarat que confiés en cap dels seus companys de feina	Me sorprende que haya venido nadie; Al bedel le irrita / molesta / agobia tener que decirle a nadie cómo se rellenan los impresos; Resultaba espantosa la idea de confiar en ninguno de sus amigos; Me pareció extrañísimo / muy extraño que Juan moviera un dedo por él; Es realmente insólito que a Luís se le ocurran semejantes tonterías; Fue una locura venir siquiera
5.24	Sense fer cap mena de treball ni d'exercici, ¿com vols aprovar?; En comptes de trucar a ningú, mira si pots solucionar el problema tu sola; Haurem d'intervenir abans que ningú demani la paraula; Busca l'arracada fins que aparegui enlloc	Antes de mover un dedo por semejante personaje piénsatelo bien; Antes de tocar nada, lávate las manos; Es mejor que te marches sin decir nada a nadie; Escuchó toda la conversación sin que palabra alguna le sorprendiese en ningún momento

FIGURE 5: Textual search.

amidst poverty'. If the argument of the locative preposition is a bare noun, the preposition *en* 'in' is chosen (Viuen en pisos 'They live in flats.'). When preceding the name *casa* 'home', Spanish selects the static preposition *en* 'in', whereas Catalan resorts to *a* 'to': Sp. Amaneció en casa '(S)he saw the new day at home' and Cat. Va pernoctar a casa '(S)he stay the night at home'.

The analysis is purposely descriptive and devoid of theoretical apparatus and specialized terminology, and focuses on establishing the contrasting use of locative prepositions in both languages. Moreover, the description included, when combined with the examples, helps the learner of Catalan as L2 to capture the basics of Catalan grammar concerning locative complements, while helping him or her to avoid one of the standard errors of Catalan L2 students: the use of preposition *en* 'in' –**Viu en Barcelona* '(S)he lives in Barcelona'– instead of preposition *a* 'to' –*Viu a Barcelona* '(S)he lives in Barcelona'.

The educational benefits of HISPACAT extend to other aspects as well. On the one hand, HISPACAT provides the students of Catalan or Spanish as a L2 with a user-oriented and custom-built comparative grammar which, unlike any currently available tool, stresses the knowledge of the mother-tongue language to appraise the complex grammatical aspects of the second language. On the other hand, the inclusion of a big number of analyzed examples of both grammatical and ungrammatical constructions which can be retrieved through a series of complex and robust query systems, helps the teacher to obtain ready-made examples of all kinds of constructions to illustrate his or her grammatical explanations in the classroom, and facilitates the development of specific teaching materials.

[4.2] *Comparative approach: syntactic interference*

A second major class of applications of the HISPACAT database concerns the creation of an empirical playground for testing hypotheses about syntactic interference in cases of bilingualism or L2 learning. Note a particularly clear example of interference reflected in the following construction (I emphasize the relevant part):

(17) quantifier cada 'each/every' in temporal expressions which apparently aren't distributive'

- EX-CAT: Cada diumenge vaig al cinema
- ANAL: [...] With temporal expressions, the distributive use is clear in *Cada dia va a escola a una hora diferent* 'Every day (s)he goes to school at a different time', but not in *Cada dia va a escola* '(S)he goes to school everyday', where Spanish resorts to the nondistributive quantifier (*Todos los días va a la escuela* '(S)he goes to school everyday'). [...]
- *COMMENT: Indeed, maybe Spanish allows such a construction (it is clearly found in the Spanish spoken by Catalan speakers).*

After the description of the differences concerning the use of this quantifier in standard Catalan and Spanish, the *comment* field includes valuable information about the interference of the Catalan system in the Spanish dialect spoken in Catalonia.

Consider now an example of interference the other way around (I emphasize the relevant part):

(18) phase adverbs *ya* 'already' and *todavía* 'still' in immediate postverbal position

- DEN-CAT: adverbis de fase *ya* i *todavía* en posició immediatament postverbal
- DEN-ESP: adverbios de fase *ya* y *todavía* en posición inmediatamente postverbal
- EX-CAT:
- REF-CAT:
- EX-ESP: Está ya/todavía aquí '(S)he is already/still here'
- REF-ESP: GCC 26.2.2.1

- ANAL: Both Spanish and Catalan allow the adverb-verb ordering: Sp. *Ya/-Todavía está aquí* '(S)he is already/still here'; Cat. *Ja/Encara és aquí* '(S)he is already/still here'. The existence of the other order in Spanish seems to be related to the existence of a higher (left) position for the verb, which one can appreciate in other constructions like Sp. *Tiene usted mucha suerte* 'You are very lucky' o Sp. *Es quizá cierto*, where the verb may precede both the subject and certain adverbs in Spanish, but not in Catalan (or at least less easily).

- *COMMENT: As it is usually the case, one finds the spurious copy of the Spanish order in Catalan, particularly in formal written texts.*

Crucially, this is the kind of linguistic information that can hardly be found in monolingual grammars or textual databases, but is best suited to contribute to help testing empirical hypotheses about grammatical symmetries and asymmetries stemming from language contact.

[5] CONCLUSIONS

From the preceding discussion we can conclude that the HISPACAT Catalan-Spanish contrastive database of constructions is linguistically and conceptually based, and grounded on the theoretical framework of syntactic microvariation. Its main theoretical goals are the identification of the "atoms" of linguistic microvariation, and offering a snapshot of those grammatical areas more vulnerable to syntactic variation and interference. From an applied point of view, HISPACAT is conceived as a major database for contrasting methods and hypothesis in bilingualism and L2 learning, and as a dynamic comparative grammar and catalog of commented examples for teachers and students of Catalan or Spanish as L2. To attempt these goals, HISPACAT develops an ontology of linguistic concepts decomposing the basic features of each particular construction, while allowing a flexible and robust system of concept queries.

REFERENCES

Baker, M. 1996. *The polysynthesis parameter*. Oxford/New York: Oxford University Press.

Borer, H. 1984. *Parametric syntax: Case studies in semitic and romance languages*. Dordrecht: Foris.

Bosque, I. & V. Demonte. 1999. *Gramática descriptiva de la lengua española*. Madrid: Espasa.

Chomsky, N. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Espinal, M. T. 2002. La negació. In J. Solà, M. R. Lloret, J. Mascaró & M. P. Saldanya (eds.), *Gramàtica del català contemporani*, vol. 3, 2727–2797. Barcelona: Empúries. Chapter sintaxi-24.

Fernández-Ramírez, S. 1951. *Gramática española*, chap. 4 El Verbo Y la Oración. Madrid: Arco/Libros.

Fábregas, A. 2007. The exhaustive lexicalisation principle. In Monika Bašić, Marina Pantcheva, Minjeong Son & Peter Svenonius (eds.), *Tromsø working papers on language & linguistics*, vol. 34, 2, 165–199. Nordlyd. Special issue on Space, Motion, and Result.

Garrido, J. 1999. Los actos de habla. las oraciones imperativas. In I. Bosque & V. Demonte (eds.), *Gramática descriptiva de la lengua española*, vol. III, 3879–3928. Espasa, Madrid.

Goldberg, A. E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Hernanz, M. L. 1999. El infinitivo. In *Gramática descriptiva de la lengua española*, vol. II, 2197–2356. Espasa, Madrid.

Jaeggli, O. & K. Safir. 1986. The null-subject parameter and parametric theory. In *The null-subject parameter*, 1–44. Dordrecht: Kluwer.

Kayne, R. S. 1996. Microparametric syntax. some introductory remarks. In J. Black & V. Montapanyane (eds.), *Microparametric syntax and dialectal variation*, ix–xvii. Amsterdam/Filadelfia: John Benjamins.

Kayne, R. S. 2000. *Parameters and universals*. New York: Oxford University Press.

Kayne, R. S. 2005. Some notes on comparative syntax, with special reference to english and french. In G. Cinque & R. S. Kayne (eds.), *Handbook of comparative syntax*, 3–69. New York: Oxford University Press.

Payrató, L. 2002. L'enunciació i la modalitat oracional. In J. Solà, M. R. Lloret, J. Mascaró & M. P. Saldanya (eds.), *Gramàtica del català contemporani*, vol. 2, 1149–1220. Barcelona: Empúries.

Quer, J. 2002. Les relacions temporals i aspectuals. In J. Solà, M. R. Lloret, J. Mascaró & M. P. Saldanya (eds.), *Gramàtica del català contemporani*, vol. 3, chap. sintaxi-25, 2799–2866. Empúries.

Ridruejo, E. 1999. Modo y modalidad. el modo en las subordinadas sus- tantivas. In I. Bosque & V. Demonte (eds.), *Gramática descriptiva de la lengua española*, vol. II, 3209–3252. Madrid: Espasa.

Rizzi, L. 1982. *Issues in italian syntax*. Dordrecht: Foris.

Saldanya, M. Pérez. 2002. Les relacions temporals i aspectuals. In J. Solà, M. R. Lloret, J. Mascaró & M. P. Saldanya (eds.), *Gramàtica del català contemporani*, vol. 3, chap. sintaxi-22, 2567–2662. Empúries.

Solà, J., M. R. Lloret, J. Mascaró & M. Pérez Saldanya (eds.). 2002. *Gramàtica del català contemporani*. Barcelona: Empúries.

AUTHOR CONTACT INFORMATION

Xavier Villalba
Universitat Autònoma de Barcelona
Dept. de Filologia Catalana
Facultat de Lletres (edifici B)
ES-08193 Cerdanyola del Vallès
Spain
Xavier.Villalba@uab.cat

OSLa J. B. Johannessen (ed.), *Language Variation Infrastructure*, Oslo Studies in Language 3(2), 2011. 147–149. (ISSN 1890-9639)
http://www.journals.uio.no/osla

AUTHORS

Patrik Bye
Patrik Bye is a researcher with the Center for Advanced Study in Theoretical Linguistics at the University of Tromsø. He has published scholarly articles on a variety of topics including the syllable structure, quantity, and stress systems of the Finno-Ugric Saami languages, North Germanic accentology, and phonologically conditioned allomorphy.

Harald Hammarström
Harald Hammarström studied Linguistics and Computer Science at the University of Uppsala. He continued on to do his PhD in Computational Linguistics at Chalmers University. His main research interests are linguistic typology, language description and computational linguistics. He now works at the Max Planck Institute for Evolutionary Anthropology in Leipzig, where he is involved in macro-level projects of linguistic diversity, language documentation and areal/genetic relations. He is also engaged in the documentation of Mor, a Papuan language of West Papua, Indonesia.

Anton Karl Ingason
Anton Karl Ingason is an MA student at the University of Iceland. His interests include variation and change in morphosyntax and at its interfaces. His area of interest includes theoretical models of productivity and rule blocking. In addition to work on theoretical linguistics he has worked on natural language processing for Icelandic and he is the author of Lemmald, a lemmatizer which is a part of the IceNLP toolkit for processing Icelandic text. He is currently a member of the team which is building the Icelandic Parsed Historical Corpus (IcePaHC) as part of the project Viable language technology beyond English - Icelandic as a test case.

Janne Bondi Johannessen
Janne Bondi Johannessen is professor of linguistics and language technology at the Text Laboratory, Department of Linguistics and Nordic Studies at the University of Oslo. She has been in charge of several projects for developing new language infrastructure, such as tagging and annotation tools, language databases, search interfaces, monolingual corpora, parallel corpora and not least speech cor-

pora, all of which are used widely in the Nordic countries and internationally. Her research publications are both within language technology and linguistics, including dialectology. She is president of NEALT, North European Association of Language Technology.

Jan Pieter Kunst

Jan Pieter Kunst is employed at the Meertens Institute, Royal Netherlands Academy of Arts and Sciences. He originally studied Dutch language and literature at the universities of Utrecht and Amsterdam, specializing in medieval Dutch literature, with Computers in the Humanities and Artificial Intelligence as secondary subjects. Since 2002 he has been employed as a software developer at the Meertens Institute, where he creates web applications (mostly written in PHP with MySQL in the backend) to make the research data and collections of the Meertens Institute available to the world.

Therese Leinonen

Therese Leinonen works as a postdoc researcher at the Society of Swedish Literature in Finland. Her main research interests lie in computational dialectology, sociophonetics and variationist linguistics in general. She studied Scandinavian languages and literature at the University of Helsinki, and has worked as researcher at the Research Institute for the Languages of Finland compiling the Dictionary of Swedish Dialects in Finland. She was affiliated with a dialectometric research project at the University of Groningen 2006-2010 and received her doctoral degree in 2010.

Sebastian Nordhoff

Sebastian Nordhoff studied General Linguistics, Computational Linguistics, and Spanish at the Universities of Cologne and Seville. His main research interests are language description and documentation and technical requirements and solutions for that. After doing fieldwork on Paraguayan Guaraní for his MA and Sri Lanka Malay for his PhD, he now works at the Max Planck Institute for Evolutionary Anthropology in Leipzig, where he is developing several projects in the area of electronic publishing. His current projects are Langdoc/Glottolog as described in this paper and the grammar authoring platform GALOES.

Eiríkur Rögnvaldsson

Eiríkur Rögnvaldsson is professor of Icelandic language and linguistics at the Faculty of Icelandic and Comparative Cultural Studies, and has held a position at the University of Iceland since 1986. His research is mainly within Icelandic syntax and he has published a number of papers on both Modern Icelandic syntax and the syntax of Old Icelandic. He has also written articles and textbooks on Modern Icelandic phonetics, phonology and morphology. In recent years, his interests have turned to corpus linguistics and language technology. He has been project leader or steering group member of several Icelandic language technology projects. He has also taken part in a number of Nordic projects and networks in this area.

Diana Santos
Diana Santos, researcher at SINTEF ICT, has worked in natural language processing of Portuguese for 25 years, having written a MsC thesis on machine translation in 1988 and a PhD on corpus-based contrastive semantics in 1996, both at Instituto Superior Técnico, Technical University of Lisbon. She leads Linguateca, an international resource center for the computational processing of Portuguese since 1998. Her main research interests are semantics, translation, evaluation, and corpus-based approaches to linguistics.

Einar Freyr Sigurðsson
Einar Freyr Sigurðsson is an MA-student at the University of Iceland. The topic of his thesis is the New Passive in Icelandic, its nature and origins. He is also working in the IcePaHC project which aims to build a parsed historical corpus of Icelandic. His research interests are syntactic change and variation and quantitative methods in linguistics.

Xavier Villalba
Xavier Villalba is professor of Catalan linguistics and chair of the Dept. de Filologia Catalana at the Universitat Autònoma de Barcelona. His main research interest include syntax and the syntax/semantics and syntax/pragmatic interfaces, on which he has published articles in Lingua, Journal of Pragmatics, Catalan Journal of Linguistics, Portuguese Journal of Linguistics, and Caplletra.

Franca Wesseling
Franca Wesseling, MA, works at the Meertens Institute, Royal Netherlands Academy of Arts and Sciences. She finished her Research Master Linguistics at the University of Amsterdam in 2007; her thesis was on grammaticalization patterns in West African and Sinitic languages. She is interested in language variation (at macro and micro level), language change, syntax, and computational linguistics. Wesseling works at the Meertens Institute within the Edisyn Project on the development of the Edisyn Search Engine, which is an online tool for making various dialect corpora interoperable.