Probabilistic tagging of minority language data: a case study using Qtag

Christopher Cox

University of Alberta

Abstract

While probabilistic methods of part-of-speech tag assignment have long received consideration in corpus and computational-linguistic research, less attention would appear to have been paid to date to the development of tagging accuracy over rounds of iterative, interactive training in applications of these methods. Understanding this aspect of probabilistic tagging is arguably of particular importance to the successful construction of minority language corpora, where financial resources for corpus development are often limited and no fixed standards for either orthography or part of speech assignment may necessarily exist. This paper therefore presents a case study in the application of pure probabilistic tagging, as represented by Qtag (Tufis and Mason, 1998), to minority-language data from Mennonite Low German (Plautdietsch). Concentrating upon the relationship of several factors (including training data size, tag set complexity, and orthographic normalization) to the development of tagging accuracy, the present study conducts computational simulations of the iterative, interactive training process to compare the interactions of these factors quantitatively over time. The study concludes with a discussion of these factors' relevance to the development of accuracy in tagging as well as of potential confounds to the application of probabilistic tagging methods to similar minority language data.

1. Introduction

Probabilistic methods in the assignment of grammatical category labels to natural language data have long represented an area of active research in computational linguistics (see Church 1988, DeRose, 1988). Such methods, while certainly not without deterministic counterparts, have arguably been of particular importance in the recent history of corpus linguistics, where their application in large-scale corpus construction projects has often met with considerable success. This has provided researchers within both computational and corpus linguistics with quantities of tagged natural language data beyond historical parallel, and thus, in part, fostered the development of quantitative methods of linguistic modelling which make active recourse to computationally assigned attributes of their primary data.

The application of probabilistic part-of-speech assignment methods has proven profitable in many such large-scale corpus construction projects, where time, linguistic data, technical expertise, and financial resources are often comparatively abundant. Perhaps less documented, however, are the challenges of applying similar techniques when one or more of these resources is limited. Such

is commonly the case in minority-language corpus construction (see McEnery and Ostler, 2000), where both linguistic and technical-financial hurdles must often be overcome. Standards for the representation of language materials, whether in the form of codified orthographies or of descriptions of grammatical categories which must be present in any systematic tagging of the language at hand, may not have been proposed, sufficiently elaborated, or uniformly adopted within the speech communities represented, thus posing potential problems for the straightforward incorporation of available linguistic resources into an internally consistent corpus. Likewise, from a technical perspective, existing computational techniques proposed for part-of-speech tagging may falter on or simply not be amenable to the typological structure of certain languages. Depending upon the degree of detail sought in tagging, a language with polysynthetic morphology, for instance, may prove a more difficult target for dictionary-based methods of tagging, which rely to varying extents upon the recurrence of word-tag pairs to achieve accuracy, than the primarily morphologically isolating or analytical languages which have featured most prominently among large-scale corpus construction projects to date. In all such cases, limited resources for corpus construction may concomitantly limit the development of language-appropriate computational techniques or linguistic standards which might be applied to the available corpus material.

Where financial and technical resources are limited, then, it would seem important to understand which factors bearing upon corpus development might be expected to produce acceptable results and minimize overall expenditure of effort, given a set of assumed goals and available resources. The present paper offers one such evaluation of the efficacy of probabilistic computational techniques in part-of-speech tagging when applied to minority language data. This assumes the form of a case study in the use of a freely available probabilistic part-of-speech tagger, Qtag (Tufis and Mason, 1998), in the annotation of a small (approximately 120,000 token) corpus of written Mennonite Low German (*Plautdietsch*). Among its other benefits, the adopted case study format presents an opportunity to consider in concrete detail several problems commonly faced in tagging minority-language data and, thus, provides a chance not only to assess the tagging procedures adopted in this particular instance but also, through post-hoc simulation of different tagging models, to discuss alternatives which may have produced comparable results with reduced expenditure of resources. In this way, the present study seeks to offer both a description and assessment of minority-language corpus development "in action" as well as methods of corpus development evaluation which might be of broader use in similar minority-language corpus development projects.

2. Constructing a corpus of *Plautdietsch*

Plautdietsch (ISO 639-3: pdt), the language of the corpus described here, is an Indo-European language of the Germanic subgroup, formerly spoken in the area

of the Vistula Delta in northern Poland. Recent estimates place the number of Plautdietsch speakers globally between 300,000 and 500,000 (see Epp, 1993: 102–3; Epp, 2002; Gordon, 2005), although substantial variation is noted in the relative vitality of the individual speech communities which comprise this estimate. Most numerous among the present-day speakers of Plautdietsch are the descendents of Dutch-Russian Mennonites, an Anabaptist Christian denomination that originated in the Protestant Reformation. As a result of the persecution, emigration, and exile of Dutch-Russian Mennonites over the course of four centuries, sizeable Plautdietsch speech communities are to be found today on four continents and in no fewer than a dozen countries. Considerable dialectal variation is observed between disparate groups of Plautdietsch speakers distinguished historically by differing patterns of emigration and divided contemporarily by both geographical distance and several centuries of mutual isolation. For further discussion of the linguistic history and characteristics of this rather exceptional member of the Germanic language group, refer to the existing literature concerning Plautdietsch. In particular, Epp (1993) presents a thorough overview of the development of Plautdietsch in historical-linguistic context, while the origins of those written varieties represented in the present corpus are discussed in Loewen and Reimer (1985).

Much as in the construction of any other corpus, the development of a corpus of written Plautdietsch must arguably consider not only features of the available linguistic material (e.g., its representativeness within the writing traditions maintained by the contributing speech communities, the varietal features it exemplifies, etc.) and aspects of its digital representation (e.g., the selection of appropriate standards for text encoding and the representation of document structure and linguistic annotations), but also the anticipated users of the final corpus and the uses to which it might be put. The necessity of such planning is further underscored, as the previous section has suggested, when resources for corpus development are comparatively limited. This planning, however, poses problems for the would-be corpus developer: precisely which features should receive immediate attention in corpus construction, and which should be set aside as areas for future development? Which subset of the potential uses and users of the finished corpus should be selected as the specific focus of short-term development? Indeed, even limiting such considerations to the task of probabilistic part-of-speech annotation, selecting which linguistic features to annotate and to what level of detail may be a less-than-trivial undertaking and have serious consequences for the ultimate success or failure of a corpus development project to meet its stated goals, even when other linguistic and technical aspects of such a project are relatively well understood.

In this case, the present corpus of written Plautdietsch, while expected to be suitable for many possible linguistic analyses and acceptable to the speech communities whose language it represents, is intended primarily for research into the syntax of verbal complementation. For the purposes of such investigation, then, it is desirable to have detailed tagging which captures each inflectional category of finite verbs—their tense, person, number, and so on. Given the degree

of dialectal variation found between individual varieties of Plautdietsch, it may also be reasonable to suspect that the varieties themselves represent potentially relevant predictors of variation in verbal syntax. Thus, dialectal attributes should also likely be represented in some form in the corpus. From a technical perspective, the inclusion of these features in the final corpus is made feasible with computational resources for corpus construction furnished largely by the Text Analysis Portal for Research (TAPoR) laboratory at the University of Alberta. The linguistic goals of this corpus development project—that is, to produce a corpus appropriate for quantitative investigations of verbal syntax—are thus relatively clear, and institutional support significantly lessens the technical and financial burdens that might otherwise discourage such an endeavour. Nevertheless, the time required to develop such a corpus is shared with that of later research into verbal syntax, and the management of this resource is important to the success of both projects. Time investment in corpus construction should therefore be minimized wherever possible.

While the benefits provided by institutional support are considerable and should not be underestimated, the development of a corpus of Plautdietsch nevertheless faces several challenges commonly encountered in minority-language corpus construction. The present discussion limits its attention to three such challenges in particular:

1. *No single orthographic standard exists for Plautdietsch*. Both the relatively short history of Plautdietsch as a written language (see Epp 1996: 3) and the geographical dispersion of Plautdietsch speakers globally have contributed to the wide range of orthographic systems and conventions attested in contemporary texts. Spelling systems may vary not only between individual authors, but even between the individual works of a single author, with several noted Plautdietsch writers having elaborated upon their own orthographic standards over the period of several decades (see Nieuweboer, 1999). Thus, Plautdietsch spelling systems may vary where the represented varieties themselves presumably do not (e.g., in the pronunciations of a single author who has developed a spelling system over time) or, in the case of phonetically close orthographies, present potentially valuable sources of information on phonological variation and speakers' perceptions thereof.

   While common conventions for the representation of certain phonemes have emerged across many such spelling systems, this fact in itself does not present an immediate remedy to many of the problems encountered with the presence of multiple orthographic standards in a single corpus. Diversity in spelling is unlikely to cause probabilistic tagging procedures to fail entirely; it is likely, however, to increase the difficulty of inductively training an effective probabilistic model with lexically conditioned probabilistic tagging systems, as the number of orthographically distinct word forms grows with each new spelling system represented in the corpus, and thus increases the overall number of types

which the tagging system must either come to recognize or learn to predict effectively. Perhaps more problematically, having multiple spelling systems represented as such in the corpus renders exhaustive search and retrieval difficult, if not impossible. For a task as basic as retrieving all instances of a given word (to compute a lexical frequency or dispersion measure, for instance), one must essentially search for each possible spelling of that word, a task that would require a priori knowledge of each spelling system in use in the corpus (while making the somewhat generous assumption that each system has been applied consistently and without significant variation in each source work). The magnitude of this problem becomes all the more apparent when attempting to search for pairs or sequences of collocates in the corpus, with a combinatorial increase in the potential number of orthographic variants which must be taken into account in each search.

In addition to the technical challenges posed by variation in spelling, the choice of orthographies remains an issue of some contention among authors of Plautdietsch. Individual writers and publications often express strong preferences for particular orthographic standards or conventions. If the final corpus is to be considered acceptable by the larger Plautdietsch speech community, the orthographies chosen by those authors whose works are represented in the corpus must likely be preserved in some form in the corpus.

2. *No corpora of Plautdietsch have been published to date*. While several studies of Plautdietsch (e.g., Klassen, 1969; Hooge, 1973) have made reference to private corpora assembled largely from independent fieldwork, excerpts of which have occasionally been published in edited form (see Klassen, 1993), no publicly available digital corpora of Plautdietsch exist. No systems of conventions for representing part-of-speech categories ('tagsets') have been proposed for this language, nor indeed is there significant consensus among existing grammatical descriptions of Plautdietsch varieties as to the grammatical categories which must be present in any adequate representation of the language. (Even relatively basic features of the language remain heavily disputed, such as the number of distinct cases in Plautdietsch for which determiners and adjectives inflect, rendering their codification in a standardized tagset more difficult.)
3. *Dialectal variation*. As was noted earlier, substantial variation exists both between and within national varieties of Plautdietsch in their lexical and morphosyntactic features. Whereas the former category of lexical variation is unlikely to prove problematic for probabilistic tagging—word forms characteristic of one or another particular variety will appear as types of limited dispersion in the corpus—the latter category of morphosyntactic variation poses potential difficulties for the training of probabilistic taggers and effective search and retrieval within the finished corpus. The reasons for this are much the same as those for problems related to

orthographic variation: morphological variation in the realizations of common inflectional features (e.g., the form of the regular nominal plural suffix *–e[n]* or of the infinitival verb suffix *–e[n]*) causes an increase in the overall type count in the corpus, presenting a greater number of unique word forms with which the probabilistic tagger must grapple and for which the corpus user must know to search. These problems are only compounded in an orthographically unnormalized corpus: searching for all occurrences of a given word in a dialectally diverse and orthographically unnormalized corpus must take into account not only all possible dialectal variants of that word, but also all possible spellings of each such variant. While perhaps feasible for certain simple lexemes, this solution quickly becomes intractable as the length of the search and the orthographic and dialectal variability of the sought-after tokens increase.

Given these challenges, then, a three-stage construction procedure was adopted for the present corpus of Plautdietsch which was intended to address each of the above challenges in turn. These stages are as follows:

1. *Orthographic normalization*. A separate version of each text in the corpus was created with the spelling normalized according to a published orthographic standard for Plautdietsch. Each orthographically normalized token in these separate versions was cross-referenced with the token or tokens to which it corresponds in the original text via a unique identifier. Thus, the corpus maintains both the original, authorial spelling of each text, as was required to respect the orthographic wishes of each author, as well as a standardized representation of the same, with both versions available for later use in linguistic inquiry.
2. *Adaptation of an existing tagset to Plautdietsch*. Rather than attempt to develop an entirely new tagset for Plautdietsch, an existing set of conventions proposed for the assignment of part-of-speech tags to Standard German, the Münster Tagset German (MT/D; Steiner, 2001) was adapted to suit Plautdietsch. Where existing grammatical descriptions were in agreement, categories in the Standard German tagset, which are not found in Plautdietsch (e.g., a distinct morphological verb form representing the subjunctive aspect, which has merged with the simple past in Plautdietsch), were eliminated. Where grammatical descriptions of Plautdietsch were in disagreement, the more detailed categories of the larger tagset were generally preserved. This process resulted in a significant reduction in the overall number of categories for annotation, leaving 99 distinct part-of-speech tags. Adapting an existing set of published tagging conventions for a related language, while clearly not an option available to all minority languages, proved to be of benefit here, allowing greater attention to be given to those particular cases in which conventions of the source tagset appeared out of step with features of the

target language than might have been otherwise possible, had a comparable tagset for Plautdietsch been developed *de novo*.

3. *Probabilistic tagging*. In order to apply the adapted tagset to the now orthographically normalized Plautdietsch texts, corpus construction employed a language-independent, pure probabilistic tagger, Qtag (Tufis and Mason, 1998). Several features of Qtag motivated its selection over other, comparable probabilistic tagging systems, not the least of which was its provision of a reference implementation of the probabilistic tagging algorithm on which it relies. As this implementation was written in Java and made freely available for non-commercial use, and provided a well-documented application programming interface (API) supporting Unicode, it was anticipated that Qtag might be integrated into the current project with minimal expenditure of resources, financial and otherwise. Since the basic Qtag algorithm has been published in Tufis and Mason (1998), it would also had been possible to reimplement this system independently in another programming language or environment, had the need arisen. While Qtag is certainly not alone in the class of probabilistic taggers offering similar features, it nevertheless presents a reasonable point of departure into probabilistic tagging.

Of these three stages of corpus construction, the final one, in which the adopted tagset was applied to the normalized texts which comprised the corpus, proved to be the most involved, even with the assistance provided by a probabilistic tagger. As Qtag requires a set of correct tag-token pairs from which to induce its initial probabilistic model, it was not possible to apply the tagger to the entire corpus immediately, and no other corpora of Plautdietsch were available from which such training data could be drawn. Instead, corpus texts were tagged with the adopted tagset incrementally in an iterative, interactive process. At the beginning of this process, each corpus text was divided into c segments, or chunks, of n tokens. For the first text in the corpus, tags were assigned manually to each of the n tokens appearing in its first chunk, producing a total of n-correct tag-token pairs. Qtag was then trained on these first n tag-token pairs, forming an initial probabilistic model of the language, consisting of both a matrix of transition probabilities between the observed sequences of tags as well as a probabilistic lexicon of observed token-tag associations. This model was then used as input to the first iteration of tagging: taking this model as an indication of how tags are meant to be applied, Qtag was made to assign tags probabilistically to the n tokens of the next chunk of text. These probabilistic assignments were then corrected manually—the interactive portion of this process—and Qtag was retrained on the collection of $2n$-verified tag-token pairs, which were then available as training data. The same process was repeated for all remaining chunks in all remaining documents, having Qtag assign tags to each chunk on the basis of the corrected examples it had been trained on thus far, and these assignments then being corrected by hand and fed into the ever-growing set

of training data from which Qtag inductively built its model of lexical and tag-sequence probabilities.

This iterative, interactive process was repeated until all chunks in the corpus had been successfully tagged and corrected. This process thus made use of the ability of Qtag to assign tags probabilistically given even a small amount of manually corrected input, and to develop progressively more informed (and, with any luck, more accurate) models of the language and the tagset under consideration as greater amounts of corrected training data became available over subsequent iterations. All told, this corpus construction process, when applied to the present corpus, resulted in approximately 120,000 tokens of orthographically normalized Plautdietsch text, tagged entirely according to the adapted tagset.

3. Modeling corpus construction

While ultimately successful, as noted in the previous section, this iterative, interactive process of tag assignment proved to be the single most time-consuming and labor-intensive segment of the larger corpus construction procedure. As such, it was also the most crucial element to the completion or abandonment of corpus development plans. Had this stage taken more time to finish than was anticipated or more resources than had been allotted to it, tag assignment may have needed to be set aside or the entire corpus development plan reconsidered. The question might therefore be asked: what could have been done to reduce the burden of corpus construction as a whole, without lessening the quality of the resulting data? Was it necessary, for example, to provide Qtag with normalized spellings in advance of tagging? Would omitting this step, itself a considerable investment of time, have reduced to any significant extent the rate at which accuracy developed in the later iterative, interactive tagging process? Should the tagset adopted for application to the corpus texts have been more or less elaborate than it was? Should greater numbers of tokens have been tagged in each iteration of the tagging process?

It is clear that the modest success achieved in the present method of corpus construction in arriving at an application of the given tagset to the corpus data cannot reasonably be taken to imply anything more than that it was possible, given the noted investment of time and effort. Open questions remain as to whether or not better solutions to this same problem might have been found, thus decreasing overall resource expenditure while achieving the same final product—or indeed, which if any of the decisions made in planning corpus development may have borne most heavily upon the investment of effort ultimately required. The answers to these questions, however, are of immediate interest, to continued corpus development within the present project and potentially to comparable minority-language corpus construction projects as well and, thus, arguably deserve further attention.

In order to begin to address questions such as these, the present study opts to conduct computational simulations of different models of interactive, iterative

tagging. In these simulations, different combinations of parameters to the tagging process are varied systematically to represent distinct combinations of choices that might have been made during corpus planning. Having the final, corrected corpus on hand, it is possible to recreate in each automated simulation the iterative, interactive tagging process for the combination of parameters under investigation. That is, having on hand the final set of corrected training data for the entire corpus, it is possible to automate the training of Qtag on successively larger portions of these data, monitoring at each stage the accuracy of its tag assignments to the next chunk of the remaining corpus data. Simulations are thus able to observe and quantify the performance of a given constellation of corpus design decisions, relative to assumed goals and requirements; they can, in effect, undertake a simple exploration of the parameter space of possible corpus design decisions, comparing the relative merits of each such option according to the criteria used for evaluation.[1]

As parameters to the simulations conducted here, three classes of corpus design decisions are considered.

1. *Orthographic normalization*. Simulations of probabilistic tagging are performed using both orthographically normalized and orthographically unnormalized data.
2. *Chunk size*. Individual simulations vary the number of tokens to which tags are assigned and subsequently corrected in each round of iterative, interactive tagging. For the purposes of this study, 14 chunk sizes are considered (i.e., 100; 200; 300; 400; 500; 750; 1,000; 1,500; 2,000; 3,000; 4,000; 5,000; 7,500; and 10,000 tokens per chunk). Although the selection of token sizes made here is somewhat arbitrary, its range is arguably not altogether unrealistic for manual tag assignment correction. However, nothing prevents the consideration of other chunk sizes as well.
3. *Choice of tagset*. Simulations vary the complexity of the tagset being applied. Two new tagsets are defined as surjections from the categories of the original 99-tag tagset to sets of categories having 50 and 13 tags, respectively. This results in three tagsets of differing complexity being compared here: where the 99-tag tagset assigns distinct labels to the possible combinations of tense, person, and number features on inflected verbs, for instance, these labels all reduce to the single category of *V* (verb) in the 13-tag tagset.

Each combination of these parameters—each model of iterative, interactive tagging—is simulated and evaluated in terms of two measures. The first is the rate of accuracy development over time. Of interest here are models in which initial accuracy is high and increases rapidly as more training data are supplied. The second measure is the estimated amount of time required to produce each model. The total time required to apply tagset *t* to a given model *M* is estimated as a function of the time required for the initial, manual tagging of

the first chunk c_1 and the subsequent correction of automatic tag assignments of varying levels of accuracy for the remaining $c - 1$ chunks:

$$(1) \qquad time_{total_t} = time_{manual_t}(c_1) + \sum_{i=2}^{C} time_{correction_t}(c_i, accuracy_t(c_i))$$

Estimates of the actual time requirement of manual tagging and correction at various error levels were gained through timed samples of these activities with each of the defined tagsets. While both of these metrics might be further refined and additional measures proposed by which to evaluate each simulation, they nevertheless provide in their present forms relatively intuitive means of assessing the models under consideration here.

4. Evaluating models of corpus annotation

For the purpose of exposition, the simulated models of part-of-speech tagging are divided into three classes, each concentrating upon one of the parameters introduced in the preceding section. We consider the effect of each level of these parameters upon the rate at which accuracy develops and the estimated time requirement observed for the simulated models while holding the levels of all other parameters constant, allowing us to compare the effects of each parameter individually. This procedure may be pursued exhaustively, exploring all possible combinations of parameter levels, or selectively, considering only those parameter combinations which are considered particularly promising by a given heuristic. The objection might reasonably be raised that the latter methodology, if applied blindly, may inadvertently obscure potential interactions between parameters—a particular combination of parameter levels may perform especially well or exceptionally poorly, and this important interaction be overlooked if the values of all other parameters except the one of interest are held constant. This is not an issue in the present simulations, where all parameter-level combinations have been simulated exhaustively and no such significant interactions noted. The present description of the results of simulation is therefore intended to reproduce the effect of each individual parameter, without suggesting that interactions between parameters might be safely disregarded in all such cases. We begin by considering the effects of orthographic normalization upon the rate of accuracy development and estimated time expenditure across simulations, followed by the same for chunk size and tagset choice.

5. Evaluating orthographic normalization

Guiding the investigation pursued in this section is the question raised previously: does orthographic normalization matter, either for the rate at which tagging accuracy develops over successive iterations or for the estimated overall time expenditure? The method of simulation adopted here offers one means of

addressing this question. Holding the choices of tagset and chunk size constant, we compare the results of simulations of tagging normalized and unnormalized data, thus isolating insofar as possible the effects of normalization, independent of the remaining decisions made in corpus planning.

Figure 1 and Figure 2 present the differences noted in accuracy development rates and estimated time requirements for the tagging of normalized and unnormalized data. A Wilcoxon signed rank test confirms a statistically significant difference between the rates of accuracy depicted in Figure 1 (n = 1,237, W = 742,550, $p < 0.0001$). While these figures concentrate upon the results of simulation for POS-99 (the original 99-tag tagset), similar, albeit less dramatic, relationships hold for POS-50 and POS-13 as well. On average, accuracy development rates are 20% lower for unnormalized data than for normalized data across all tagsets. As might be expected, these decreased accuracy rates correlate with increased estimated time requirements: the time required to tag the entire unnormalized corpus is estimated on the basis of these simulations to take on average 26 hours longer for POS-99, 15 hours longer for POS-50, and 11 hours longer for POS-13 than it would to tag the same corpus in orthographically normalized form. This suggests a considerable difference in both the rate of accuracy development and overall time requirements between the tagging of normalized and unnormalized text in this corpus. It would seem that, in the present corpus at least, orthographic normalization has a substantial effect upon both probabilistic tagging accuracy and estimated total time requirements—an effect observed across all tagsets and chunk sizes considered here.

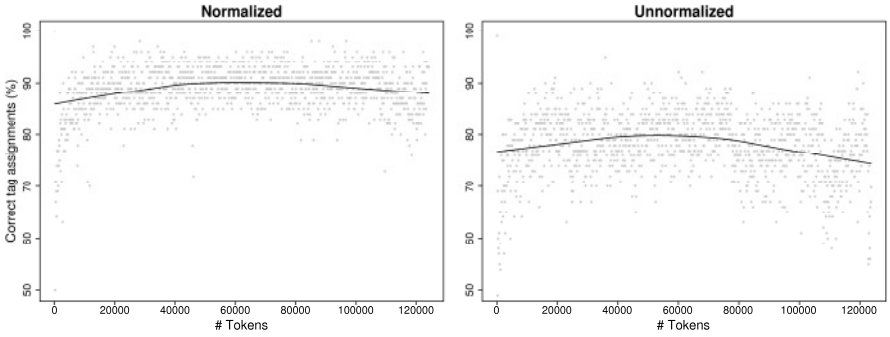

Figure 1. Rate of accuracy development observed in simulations with normalized and unnormalized data (POS-99, chunk size 100), with curves fitted by Lowess smoothers.

These results in turn raise an interesting question: what aspects of orthographically unnormalized text pose the greatest problems for the adopted methods of probabilistic tagging? While a thorough investigation of this question falls largely outside of the scope of the present investigation, it might be hypothesized that the increased number of orthographic variants found in the unnormalized corpus, serving to inflate the number of unique word forms (i.e., types) with which the probabilistic tagger must come to terms, in part causes

overall tagging accuracy to decrease. A simple negative correlation between type count and tagging accuracy would seem unlikely, however. Rather, the frequency of occurrence and consistency of tagging of individual types within the corpus, as well as the predictability of the part-of-speech categories of these types from the contexts in which they appear, may also be relevant to tagging accuracy, rendering this a less-than-trivial problem to explore, although one of potential relevance to corpus construction. Importantly, these hypothesized predictors of tagging accuracy are themselves quantitative measures derivable from corpora and tagging procedures, and thus open questions such as this to further quantitative investigation.

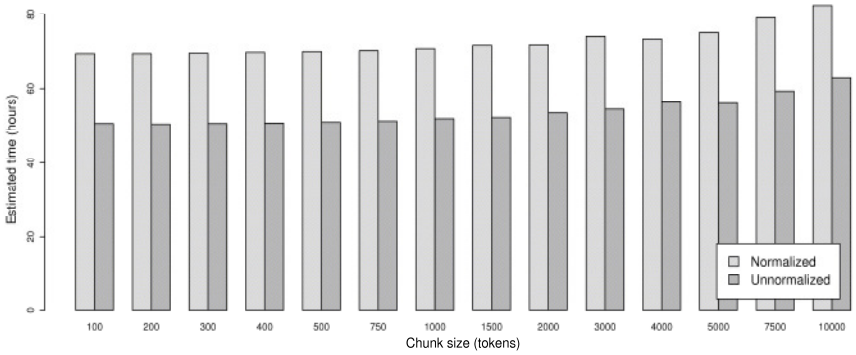

Figure 2. Estimated time expenditure for tagging normalized and unnormalized data across all specified chunk sizes (POS-99).

5.1 Evaluating chunk size

A procedure similar to the one used to assess orthographic normalization in the preceding section is adopted here to evaluate the effects of chunk size on the development of accuracy rates in tagging and the estimated overall time requirement of corpus construction. In this case, holding both the choice of tagset and orthographic normalization constant, simulations of tagging using different chunk sizes are conducted and their results compared. These results are presented graphically in Figure 3 and Figure 4. Since many chunk sizes result in similar accuracy rate measures, these rates are presented as histograms with three points for each chunk size (i.e., tagging accuracy rates at the beginning, middle, and end of tagging the corpus).

Inspection of these figures suggests an immediate difference between the relationship of chunk size and that of normalization to the assumed measures of accuracy development and estimated time investment. First, statistical investigation using Pearson's product moment correlation test suggests a general correlation between the rate of accuracy development and chunk size in initial (r = 0.6398, df = 12, p = 0.01373) and final (r = 0.9451, df = 12, $p < 0.000001$) stages of tagging, albeit not significantly in medial stages (r = -0.4887, df = 12, p

= 0.07617). Likewise, positive correlations between estimated time requirement and chunk size are noted for POS-99 (r = 0.9939, df = 12, p < 0.000001), POS-50 (r = 0.9970, df = 12, p < 0.000001), and POS-13 (r = 0.9976, df = 12, p < 0.000001).

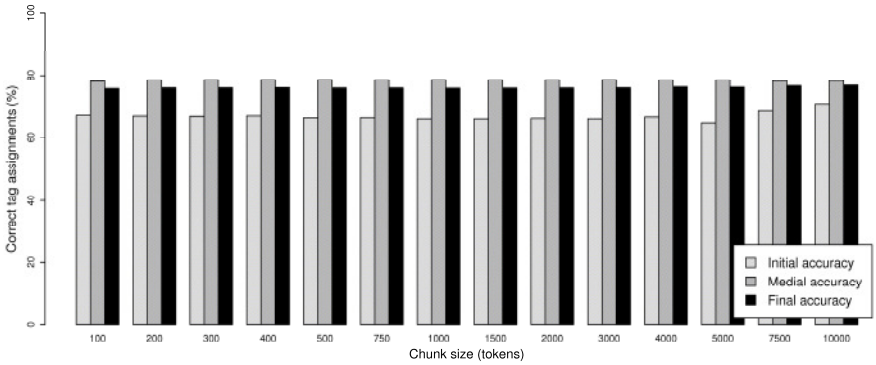

Figure 3. Rates of accuracy development observed in the initial, medial, and final stages of simulations of tagging across all specified chunk sizes (POS-99, normalized data).

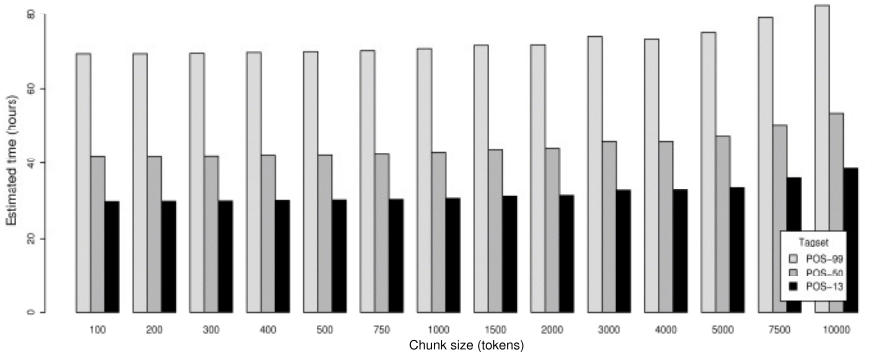

Figure 4. Estimated time expenditure of tagging in each tagset across all specified chunk sizes (normalized data).

While accuracy rates remain essentially the same for all chunk sizes less than or equal to 5,000 tokens, considerable differences are found in estimated time requirements, with smaller chunk sizes (roughly, those of less than 2,000 tokens) consistently taking less time than larger ones. The reason for this difference lies in the amount of time required to assign tags manually to the first chunk of corpus text: without the aid of automatically assigned tags (even incorrect ones), this stage of iterative, interactive tagging typically takes longer than later stages of correction. As the size of the first chunk increases, so too does the amount of time required to assign tags to each token in that chunk by hand, thus gradually coming to outweigh any potential benefits to overall accuracy that

might have accompanied providing the tagger with a greater amount of initial training data. In the present corpus, then, it would appear sensible to attempt to minimize the amount of time required to tag the first chunk by hand and to concentrate instead upon the correction of probabilistically assigned tags in the remainder of the corpus.

5.2 Evaluating tagset choice

To evaluate the contribution of tagset choice to the rate of accuracy development and estimated overall time expenditure, all other parameters are once again held constant, and the results of simulations of tagging with each of the three proposed tagsets are compared. These results are presented in summarized form in Figure 5.

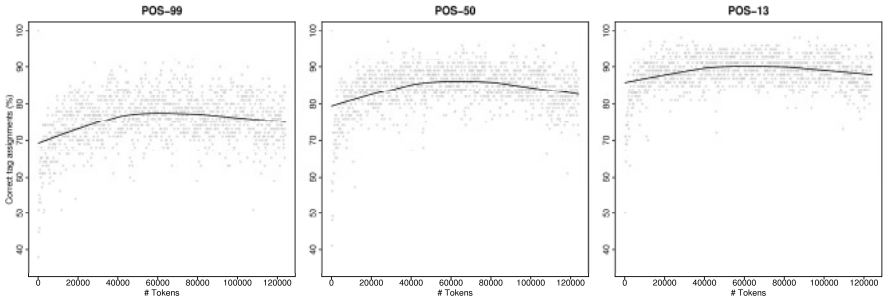

Figure 5. Rates of accuracy development across tagsets (normalized data, chunk size 100), with curves fitted by Lowess smoothers.

Much as was the case with the figures presented in preceding sections, Figure 5 suggests a clear difference between the three tagsets and brings to the fore a general negative correlation between tagset size and mean tagging accuracy which achieves near statistical significance here (r = -0.9956, df = 1, p = 0.0595). In their mean rates of accuracy development, an average 15% improvement is found for the minimal tagset, POS-13, over the full tagset, POS-99, regardless of whether or not the texts being tagged were normalized. Likewise, the estimated time requirement for applying POS-99 to normalized data, 80.5 hours, is more than double the 36.5 hours estimated to be needed to apply POS-13 to the same data. The choice of tagsets would thus appear to represent an important factor in the overall investment of effort required to achieve full tagging, even for a corpus of this size.

While these results are intriguing and may be of use in the present corpus development project, they should perhaps be interpreted with a degree of circumspection. The alternative tagsets considered here are direct adaptations of the full 99-tag tagset modified for use with Plautdietsch. It may be the case, however, that other, more varied tagsets may have fared better or worse when applied to the same corpus data, although this cannot easily be tested without some means of applying these tagsets to at least a sample of the present corpus

for comparison. Although the relatively simpler tagsets consistently achieved lower estimated time requirements and higher rates of accuracy development in simulations, the larger question nevertheless remains: what features of these simpler tagsets, beyond the restricted range of those labels they provide, might be cited to explain their respective degrees of success in tagging these data? What features of these or other tagsets make them more or less well-suited to the data and the probabilistic tagging system at hand? One might expect some degree of increase in accuracy by chance alone: all other things being equal, a smaller tagset provides fewer opportunities than a larger one for a probabilistic tagger to guess incorrectly, and thus might be anticipated to deliver, on the whole, more accurate responses. As in the case of orthographic normalization, however, a thorough answer to questions such as these pertaining to tagset design lies outside the purview of this study, which is concerned primarily with tagset application. Nevertheless, methods of evaluation such as the simulation-based techniques employed here might be of some use in developing quantitative measures of relative tagset complexity, and thus be helpful in addressing these open issues in corpus design as well.

6. Summary and conclusions

In the present case, evaluation of the results of simulation would appear to suggest the following guidelines as relevant to successful tagging:

1. With regard to orthographic normalization, improvements in the rate of tagging accuracy development may be substantial when working with normalized data (here, on the order of 20% more accurate). However, these gains must be weighed against the cost of normalization itself. Particularly in the case of minority languages, where no single orthographic standard (or appropriate spell-checking software, for that matter, if the process of orthographic normalization is to be partially automated) may necessarily exist, achieving consistent normalization may itself represent a considerable investment of effort.
2. When deciding on the size of chunks of corpus text to process in each iteration, smaller chunk sizes should be favored over larger ones. This recommendation is motivated by the observation that manual tagging may, in many cases, prove more time-consuming than correction of automatic tag assignments, regardless of the latter's accuracy. Relying on the probabilistic tagger early in corpus development to perform as much tag assignment as possible limits the amount of manual tagging required.
3. Less elaborate tagsets should be favored wherever corpus goals permit. While substantial gains in accuracy development rates and decreases in estimated time requirements were noted with the less detailed tagsets considered here, this observation should be interpreted with care, since no quantitative measures of tagset complexity have been established here.

> Any decisions to change the adopted tagset should be evaluated with attention to the requirements of the anticipated uses and users of the corpus: there is little benefit in applying a tagset which is too simple to be of use to those working with the finished corpus.

Such suggestions, however, must ultimately be measured not only against the quantitative estimates of accuracy development rates and overall time investment, but also against the qualitative requirements, available resources, and stated goals of the given corpus project. In this instance, where verbal complementation represents a primary focus of research, detailed coding for the inflectional features of Plautdietsch verbal morphology is needed. The choice of tagsets, then, is constrained to some extent by this requirement, although the cost of adopting a more complex tagset can be mitigated in part through orthographic normalization and the selection of a small chunk size, as the preceding simulations have suggested. Likewise, while a simpler tagset may have rendered it technically feasible to process orthographically unnormalized data, and thus avoid investment of resources in orthographic normalization, the goal of supporting exhaustive searches of the resulting corpus motivated this additional expenditure of time and effort. In short, quantitative measures cannot afford to be the sole means of assessing the relative merits of alternative corpus development strategies, although their application may indeed be of considerable benefit in corpus planning. Such measures form one important aspect of informed corpus development which exists within a larger rubric of goals and requirements. When taken together, these factors may suggest paths by which corpus construction can be guided to satisfactory completion from both quantitative and qualitative perspectives.

It is readily conceded here that determining the interactions of all such factors, whether quantitative or qualitative, in their relationship to tagging accuracy is likely impossible during corpus planning. Careful planning may well be able to anticipate many of the factors and their interactions relevant to the completion of corpus development, but the goals, requirements, and resources available to corpus construction are likely to change with the corpus as development proceeds. It is maintained here, however, that such planning and evaluation might still profitably enter into corpus construction as a regular part of the larger development process and to no less effect in the construction of minority language corpora. In the initial phases of corpus planning, consideration of reported results and guidelines proposed on the basis of other corpus construction projects, such as those put forward in this case study, might inform corpus development and suggest potential pitfalls which should be avoided. Introducing periodic evaluation, whether using simulation or other means, as an additional part of the iterative tagging process may further serve to identify problems during corpus development and present opportunities to make midstream changes as necessary. Even if all relevant interactions are not apparent in advance of corpus implementation, this would not seem to imply that corpus

development cannot benefit from previous experience in corpus design or from regular evaluation during corpus construction.

The selection of pure probabilistic methods over other methods available for tag assignment is also far from given. This decision, too, might be informed by consideration not only of the technical requirements of corpus development and the requirements implied by corpus end-use goals, but also the typological features of the language and the characteristics of the available sources of data, as noted previously. If one of the goals of corpus development is to integrate corpus documents with other available sources of linguistic data (e.g., existing digital dictionaries, word lists, collections of morphological parses, etc.), this may encourage the use of hybrid tools which permit concurrent lemmatization or other annotation. Whereas the present task appears well suited to the use of Qtag, benefitting from this tool's availability and simple integration into larger projects, this should not be taken to suggest that other, comparable tools might not be appropriate in the same or similar contexts or that their application to corpus development might not benefit from the processes of evaluation discussed here as well.

Computer-assisted methods of part-of-speech assignment present a range of complex technical and practical problems for the construction of modern corpora. As a stage of corpus development during which considerable resources must commonly be invested, corpus tagging represents a particular challenge for minority-language corpus development, where such resources are often limited, and thus an area in which quantitative, computational methods of evaluation might be of use in corpus development planning. While computational methods of tagging and evaluation are of clear importance to the progress of many corpus development projects, and thus arguably merit the attention which has been given to them here, it has been insisted here they arguably cannot afford to be the sole object of inquiry. Rather, consideration is also required of the larger context in which such methods are applied and of the resources, research requirements, and (socio)linguistic conditions which bear upon corpus construction as a whole.

Case studies of minority-language corpus construction present one means of contributing to an understanding of such problems in context. The results of such case studies, which are in most instances language and corpus-specific, might serve to offer general direction for further quantitative studies of corpus and tagset design, as several sections of this study have suggested. By the same token, case studies might offer an honest assessment of the challenges facing corpus construction and corpus-based language documentation in the use of contemporary computational techniques, providing guidelines from which similar projects might benefit. In bringing attention to practical issues encountered in the application of current computational methods to data from underrepresented languages, evaluations of minority-language corpus construction might thus serve a twofold purpose—at once presenting computational-linguistic research with additional real-world benchmarks by which their success under varied linguistic and sociolinguistic conditions might be assessed, while fostering through the

description of current corpus construction techniques the continued development of annotated corpora for a greater number of the world's languages.

7. Notes

1. Since the number of parameters to these models of corpus construction is limited in this case, the combinatorial space which these options form can be explored exhaustively without significant difficulty (in part because simulations may be conducted in parallel, being computationally independent of one another) and all possible models thus compared with an even hand. This may not always be the case, however, since the number of possible models increases essentially exponentially in the number of parameters under consideration. With parameter-rich simulations, then, other methods of estimating the relationships of individual parameters to measures of interest may be required.

References

Church, K. W. (1988), "A stochastic parts program and noun phrase parser for unrestricted text," in: *Proceedings of the 2nd conference on applied natural language processing*. ACM, 136–143. Online at: <http://www.aclweb.org/anthology-new/A/A88/A88-1019.pdf>.

DeRose, S. J. (1988), "Grammatical category disambiguation by statistical optimization," *Computational linguistics*, 14(1): 31–39.

Epp, R. (1993), *The history of Low German and Plautdietsch: tracing a language across the globe*. Hillsboro, KS: The Reader's Press.

Epp, R. (1996), *The spelling of Low German and Plautdietsch*. Hillsboro, KS: The Reader's Press.

Epp, R. (2002), *De Jeschicht von Plautdietsch* [The history of Plautdietsch]. Kelowna, BC: unpublished m.s.

Gordon, R. G., Jr. (2005), "Plautdietsch," *Ethnologue: languages of the world, 15th edition*. Dallas, TX: SIL International. Online at: <http://www.ethnologue.com>.

Hooge, D. J. (1973), "Das Verb in der Parataxe und Hypotaxe statistisch gesehen [The verb in parataxis and hypotaxis, from a statistical perspective]," *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 26: 328–341.

Klassen, H. N. (1969), "Untersuchungen zum grammatischen Bau der niederdeutschen Mundart im Gebiet Orenburg (RSFSR) [Investigations of the grammatical structure of the Low German dialect in the region of Orenburg (USSR)]," *Wissenschaftliche Zeitschrift der Martin-Luther-Universität Halle-Wittenberg, Gesellschafts- und Sprachwissenschaftliche Reihe*, 18(6): 27–48.

Klassen, H. N. (1993), *Mundart und plautdietsche Jeschichte. Ut dem Orenburgschen en ut dem Memritjschen (Rußland)* [Dialect and Plautdietsch stories. From Orenburg and Memrik (Russia)]. Marburg: N. G. Elwert Verlag.

Loewen, H.. and A. Reimer (1985), "Origins and literary development of Canadian-Mennonite Low German," *Mennonite quarterly review*, 59: 179–186.

McEnery, T., and N. Ostler. (2000), "A new agenda for corpus linguistics—working with all of the world's languages," *Literary and linguistic computing*, 15(4): 403–420.

Nieuweboer, R. (1999), *The Altai dialect of Plautdiitsch: West-Siberian Mennonite Low German*. Munich/Newcastle: Lincom Europa.

Steiner, P. (2003), *Das revidierte Münsteraner Tagset Deutsch (MT/D). Beschreibung, Anwendung, Beispiele und Problemfälle* [The revised Münster Tagset for German (MT/D). description, application, examples, and problematic cases]. Online version: <http://xlex.uni-muenster.de/Portal/MTPD/tagsetDescriptionDE.ps>.

Tufis, D., and O. Mason (1998), "Tagging Romanian texts: a case study for QTAG, a language independent probabilistic tagger," in: *Proceedings of the 1st international conference on language resources and evaluation*. ELRA, 589–596. Online at: <http://www.racai.ro/~tufis/papers/Tufis-Mason-LREC1998.pdf>.