Camilla Balsamo* and Barbara Hans-Bianchi**

"THE PAGE" - BUILDING A PENNSYLVANIA GERMAN THESAURUS THROUGH THE CORRECTION OF OCR ERRORS

*"The PAGe" - A Digital **The**saurus of **P**ennsylvani**a** **Ge**rman* is a project of Digital Humanities carried out by an interdisciplinary research group at the University of L'Aquila, Italy.[1] Its overall aim is the creation of a digital thesaurus of the Pennsylvania German language, based on different and heterogeneous linguistic resources, dating from different periods. Due to its long contact with American English and to the lack of a received standard, this idiom presents features that opened quite of a linguistic challenge.

1. The Pennsylvania German language

Pennsylvania German (PG), or Pennsylvania Dutch,[2] is a minority language used today by more than 250.000 people in North America.[3] It evolved, through a leveling and mixing process (Keiser 2012:9), from various Middle and High German dialects spoken by the German immigrants who settled in Pennsylvania from the late 17th century until the Revolutionary War.[4] The very survival of this language after over three centuries from the first German settlement in 1683 is quite astonishing, as immigrants mostly shift to the majority language in the lapse of three generations.[5] The quite exceptional persistence of this language certainly has a strong linkage to the economic and sociocultural contexts in which the speakers have been living over time.[6]

Initially, this idiom was considered a dialect, as it was used for oral communication only, while Standard German was considered adequate for writing and formal communication. In the New World, however, Standard German inevitably lost its role as the written reference language of the community, and English came to slowly take over those formal communication functions (Van Pottelberge 2004, 298; Werner 1996, 27). Conversely, the dwindling of the linguistically related Standard language may have contributed to opening up the possibility for PG to enter the written medium and be perceived as a language in its own right (Hans-Bianchi 2016, 46).

Being born as an American language, PG has always been coexisting with the American English varieties spoken and written in the same area, and manifold influences and interferences from one language to the

* *Barbara Hans-Bianchi is Associate Professor in German Language and Linguistics at the University of L'Aquila, Italy. Her main research interests include the investigation of language contact and language change, with a particular focus on 'Pennsylvania Deitsch,' as well as studies on (the acquisition of) literacy and orthography.*

** *Camilla Balsamo is Research Assistant at the Department of Human Studies, University of L'Aquila, Italy. Her current research interests include tailored and linguistics-based OCR correction (design of algorithms meant to correct recognition results through the Python3 programming language) and the exploration of new methods for the analysis of corpora, led in a translation-related and intercultural perspective.*

[1] Research staff: Barbara Hans-Bianchi (Principal Investigator), Giovanni De Gasperis (IT Researcher), Maria Giovanna Fusco (Research Consultant), Camilla Balsamo (Research Fellow); we are grateful for the helpful collaboration offered by Mark Louden and Michael Werner.

[2] Keiser (2012, 1) argues in favor of the native term 'Deitsch' (or 'Deitsh') because of the "current broad geographical distribution of the language." The ISO 639-3 language code is pdc.

[3] This number is given by Keiser (2012, 1), but the estimates differ hugely in different sources. The ethnologue website, for example, indicates a total amount of less than 150.000 speakers (https://www.ethnologue.com/language/ pdc, last visited October 3, 2018).

[4] During the 17th century the German immigrants were very few, whereas during the 18th century Germans settled in America in huge groups, especially in a short period around 1750 (Louden 2016, 64). German immigrants who went to the US after the Independence, during the 19th century, did not integrate in this early group who labelled them "Deitschlenner" ("people from Germany"; see Louden 2016,3 ff.). For a detailed description of the historical evolution of the language and the speaker communities, see Louden 2016.

[5] See Fishman 1964; an interesting new model of language shift and maintenance is proposed in Villa and Rivera-Mills 2009.

[6] Among the relevant factors of language maintenance, there are contiguous settlement areas, limited geographic and social mobility, rural agricultural communities, strong group identity (see Louden 2016, 179).

other have been the tangible effect of language contact from the very beginning. An increasing convergence with English can be observed on different levels as a result of long-term bilingualism (Hans-Bianchi 2013, 212; Stolberg 2015, 282 ff). Eventually, the social pressure put on the language minority has been so strong as to produce a language shift to English monolingualism (Stolberg 2015, 284-285).

Those who still maintain the language as native speakers are members of sectarian communities like the Old Order Mennonites and Amish, whose strong cultural and religious identity somehow sets them apart from mainstream American society. In these settings, the language is still thriving and the number of speakers is growing.[7] This variety is often called 'Plain PG.'

On the other hand, non-sectarian (non-plain) native speakers (who in former times were the vast majority) are now in their 80s-90s. Younger, non-sectarian, people increasingly join PG classes[8] in order to learn in the classroom the language their grandparents chose not to hand down to future generations.[9]

Since the colonial period in which the language has developed in the Southeastern area of Pennsylvania, the community of speakers is nowadays located in a non-contiguous space over several US States and Canada.[10]

According to Keiser (2012, 1), the "demographic center of gravity" has shifted to the Midwest, where a new regional variety can be identified: Midwest PG (MPG), in opposition to Pennsylvanian PG (PPG) (Keiser 2012).[11]

2. Codification and standardization

To the present day, PG has not developed a standard variety. The two aforementioned regional varieties, PPG and MPG, are mutually intelligible to a very high degree (Keiser 2012, 1). There are more fine-grained local differences within each main regional variety, which at least partially go back to the different origins of the German settlers (see Seifert 2001). The sociolinguistic varieties of Plain and Non-Plain PG are largely overlapping with the Pennsylvania vs. Midwest distinction,[12] and there is no apparent difference in prestige of any one variety over the others.[13]

Importantly, PG has remained a mainly spoken language, although it has been publicly used in writing ever since the 1840s-1850s (Werner 1996, 44), first in newspaper columns, poems and short prose texts, with an increasing frequency from the 1860s and 1870s on (see Louden 2016). Nowadays, the internet gives access not only to written texts but also to audio and video files in spoken PG.[14]

The lack of institutions that could represent the whole language community and thus officially back a standard variety, makes it vain to look for a "core codex" (Klein 2014, 224). Nevertheless, we do find writings explicitly aiming at codification and standardization which enjoy at least partial acceptance in the language community; as such, they can be considered as part of a "para-codex" (Klein 2014, 224).

One of the central pieces of this para-codex is the PG lexicography.[15] The first attempts to systematically present the PG vocabulary date back to the 1870s[16], and until today there have been numerous publications

[7] Louden (2006, 90); *Ethnologue* assigns to PG the EGIDS (Expanded Graded Intergenerational Disruption Scale) grade 5 'Developing:' "The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable."

[8] See https://hiwwewiedriwwe.wordpress.com/learn-the-dialect (last visited October 3, 2018).

[9] The most important reasons for this language shift are given, among others, in Stolberg 2015, 284-285.

[10] The distribution of PG in the United States is shown in the following map, based on Census 2000: https://en.wikipedia.org/wiki/Pennsylvania_German_language#/media/File:Pennsylvania_German_distribution.png (last visited October 3, 2018).

[11] Keiser's investigation focuses on the phonological (and lexical) differences, but there are several structural divergences too. See Fuller 1996.

[12] In the Midwest, we only find PG Plain speakers, whereas in Pennsylvania there are Plain and Non-Plain speakers.

[13] For the issue of diasystematic variability in PG, see Hans-Bianchi 2016, 44-46.

[14] Here are just two examples of widely used internet sites in PG: https://hiwwewiedriwwe.wordpress.com; https://www.jw.org/pdc.

[15] Descriptions of the PG grammatical system are less numerous, the most important being Frey 1942; Buffington and Barba 1954; Haag 1982. They all deal with the (non-plain) PPG variety, which today is no longer spoken in everyday life. For more details on PG lexicography, see Hans-Bianchi 2016.

[16] Horne 1875 (ca. 5500 items); Rauch 1879 (5000 items).

of bilingual PG/EN word lists and dictionaries, the most comprehensive one being Beam and Brown / Trout.[17] Until now, however, no monolingual PG dictionary has ever been published nor planned. In our research context, one of the most challenging aspects is the lack of a unique standardized spelling system, even in recent years. Over time, PG writers have tried many different approaches, proposing more or less systematic and coherent spelling rules (Hans-Bianchi 2016), and in the last decades, mainly two concurrent orthographic systems continue to be used, the first preferred by Pennsylvanian non-sectarians (the Buffington-Barba-Beam spelling system), the second by sectarians, mostly in the Midwestern States (the Wycliffe spelling system).[18]
To give an idea of the issue, here is an example taken from the two (almost identical) introductions of the PG Wikipedia (Example 1 and 2):[19]

Wilkum zu der Wikipedia in Deitsch! Alliebber - even yeder Leser - kann ennich Ardickele verbessere, Mischteeks tscheensche, Ardickele vergreesere, odder Ardickele schenner mache. Kannscht yuscht rum gucke aa.

Example 1: Buffington-Barba-Beam spelling system

Vilkum zu ‘s Wikipedeli in Deitsh! Alliebbah - even yaydah laysah - kann ennich ardikkela fabessahra, mishtayks tshaynsha, ardikkela fagraysahra, addah ardikkela shennah gooka macha. Kannsht yusht room gukka aw

Example 2: Wycliffe spelling system

3. The project

Starting with lexicographic texts (see section 4), other sources are being studied with the aim of integrating frequent loans from English (which are often willingly left aside by lexicographers).[20]
As we have seen in section 2, PG is rich in variants as much on the lexical as on the orthographic level.[21] We do not want to ignore these variants nor to select one form as the “best” over the others, we rather are documenting this richness linking the variants and synonyms with each other, and relating the numerous variants to their English meaning which functions as a *tertium comparationis*.
Additional lexical-grammatical information will be integrated, in order to make this digital thesaurus a useful and reliable tool for future research in PG studies. Adopting internationally used labelling systems (e.g. TEI for dictionaries, see paragraph 7) we intend to guarantee the possibility of future use and integration by other scholars.

4. Research methodology

After an initial identification and review of the main lexicographic works in PG (Hans-Bianchi 2016), the research has been carried out on two of them so far: in a first trial phase on Stine 1996 (which is copyrighted to the current days) and then on the more articulate Lambert 1924.[22] The dictionaries have been transcribed through the use of an *Open Source* software tool; an optical character recognition tool[23] applied to high

[17] We estimate that this dictionary contains more than 20000 entries. It follows the Buffington-Barba-Beam spelling system. There is no comparable dictionary for PG adopting the Wycliffe orthography.
[18] The Buffington-Barba-Beam system has been adopted among others by the *Pennsylvania German Society*. The Wycliffe spelling system was elaborated for the Bible translation directed by Hank Hershberger. For more details about the two orthographies, see Hans-Bianchi 2014.
[19] Word for word translation (BHB): “Welcome to the Wikipedia in Deitsch! Everyone – even every reader – can improve any articles, change mistakes, make articles longer, or make articles nicer. You can also just look around.”
[20] We thank Michael Werner for the permission to use the corpus materials of his Doctoral thesis.
[21] In addition to the two main spelling systems, earlier PG authors use a variety of spelling conventions, making it difficult to search for a word, if its written form does not match the spelling used in the dictionary.
[22] We have also occasionally taken into account the digitized and adapted lexicon by Peter Zacharias and his collaborators, an impressive effort carried out outside academia witnessing the thriving interest towards PG and whose authors we wish to acknowledge here.
[23] Tesseract Open Source OCR Engine http://github.com/tesseract-ocr/tesseract (last visited October 8, 2018).

resolution scans,[24] according to the standards reported in literature (see Springmann 2015; Bloomberg 1999). A special 'customization' of the "types inventory"[25] detected by the automatic recognition system allowed us to adapt the OCR to both the font and the orthographic features of each specific source-text taken into analysis.
Repeated errors occur, in this conversion phase, due to speckles and graphic imperfections of the original text, but many others are just randomly generated by the digitized encoding process – also due to the graphic format of the characters, not entirely matched by the OCR alphabet, so the character recognition algorithm will make any unexpected characters or other printed graphic forms become one of the known character in the defined alphabet.
Errors might be of different kinds: non-word detection, erroneous word-boundary detection, errors in punctuation detection, lack of diacritical marks in the output-font, tokenization errors and misrecognition of part-of-speech (POS). Some of the mistakes already existed in the .pdf (scans) of the originals we were trying to reproduce digitally. To fix the problem, humans can, of course, manually review and correct the OCR output text by hand. But human interpretation task is extremely time consuming and error-prone. So the research team has established a second goal, to be achieved together with and by means of our linguistic work on PG, namely a substantial improvement of OCR outputs by way of several Python programs that working on inputs can automatically fix errors without causing further issues.

5. Relevance of the project: Reproducibility of the model

The project "The PAGe" has invested great energies of human interpretation to study, identify and structure linguistic rules that – once they are applied to the corpus in a strictly logical and consequential order – could comply with the need to produce a 'clean' output. The major contribution of this part of the project was the development of ad-hoc programs in Python environment, meant for the reduction of a number of errors recorded during the automatic transcription phase. The goal has been and remains to perform as many corrections as possible using the automated algorithms, thus working on inputs that can automatically correct errors with no risk of generating new ones.
The correction program in Python, during the elaboration, has shown to be reliable in solving numerous systematic errors of recognition, and to be able to greatly simplify the process of systematization both of single entries and compounds (see Figure 6), with all related definitions, classifications, translations, synonyms, equivalents in German, etymologies and whatever else included within the definition itself.
This portion of the research has an intensely theoretical value. The structuring of an 'intelligent' automatic correction model, based on linguistics studies, is coherently and highly reproducible on other sources and other languages. "The PAGe" project is therefore conducting a specific study, which also has the value of a pilot study, in developing reproducible processes. The operations of correction and classification will be replicable by adapting the conceptual model to the conversion-bound corpus.

6. Approaching the text source

The research was led on different corpora. The features (style of the entry, composition of the item and information provided for the different sorts of entries) entailed the need of considering different approaches for each text source.
For what pertains to the OCR scan led on texts, factors of complexity lie in:

- presence/absence of diacritical marks
- presence of spaces and/or cues coming from punctuation
- number of characters or symbols used in the source
- predictability of the structure of the entry
- choice of the author to include variants, and collocation within the entry line/lines
- choice of the author to include or exclude compounds of the headword within the entry

[24] A 1200 DPI multi-function flat bed printer/scanner.
[25] In Tesseract the list of characters expected in the document alphabet can be specified in a configuration file. Our choice: AÄBDEFGHIJKLMNOÖPQRSTUVWXZaäæąbcdefghijklmnoöpqrstuüvwxyz~()[]"'?!,;.|-=+ <0123456789

- diversity in the composition of the definition, alternation of languages and references within the same line
- morphemic subdivision of words through the use of symbols such as slash (for prefixes and/or compounds)

Of course, the challenge degree in analyzing and digitizing a source can vary significantly. Let's consider Stine (1996) first: this source displays quite an essential original style: see Figure 1, <ausgucke>. Entry in Deitsch, comma, part-of-speech (POS: verb), semicolon, inflected form (if it is a verb, and use for reflexive if existing), no punctuation, English translation, no punctuation, sometimes notes. The spacing is very limited and sometimes, seemingly random (see Figure 2, <aa/geh>). When the lemma has more than one definition, the definition options are preceded by a number, and if the verb has a prefix, this is marked by the symbol </>

ausgucke, v.; ausgeguckt, pp. (sich die aage --) look until one is tired (in expecting some one or in staring)

Figure 2

aa/geh, v.; aagange, pp. 1.begin; 2.concern; 3.rave; 4.take fire

Figure 1

On the other hand, a less essential text, printed in an old-fashioned font, can multiply the criticalities on the typographical side, see Lambert, 1924, shown in Figure 3, <als>, for the intense presence of dashes and semicolons, but is also able to provide more helpful guidelines, due to its rigid logical structures and its punctuation coherence. In Figure 4, for the noun <babbe>, we are given an entry in two slightly different variants, accent rules included in brackets, gender, translation or explanation in English, and German reference, marked by <G>. Nonetheless, often, the unpredictability of the very structure of the entry – that could easily confuse even a human being who is conversant with the language – is almost unsolvable in a machine-led process.

als[˘], always, still, continue(d) to, in the habit of, be accustomed to, used to; while; than, as, but; — noch, still, yet; — un —, on and on. G als.

Figure 4

babbe, baba[˙˘], m, papa. G Papa.

Figure 3

In the case of Figure 5, for the verb <baehe>, Lambert provides the accent rules included in brackets, the part-of-speech (POS) and inflected form <pp. gebaeht>, followed by a micro-context in Deitsch which could be also the mentioning of some particular use, or example. The alternation of languages (Deitsch and English) within the same line is quite hard to detect for the machine.

baehe[˙˘], pp gebaeht, to toast, warm; gebaeht brot, toast. G bähen.

Figure 5

Yet compound words are even more complicated (different lexical classes can be included within the same entry, without specifications):

all || gebreichlich[˙˘ ¯˘], universal, customary; =**gemee˜**, =**gemei˜**, =**gemein**[˙˘ ¯], common(ly), universal; =**iwwer**[˙˘ ˘], everywhere, in all parts; =**mechtich**[˘˙ ˘], almighty, very; =**menanner**[˘ ˘˙ ˘], all, altogether; =**wissend**[˙˘ ˘], omniscient; =**zamme**[˘˙ ˘], altogether, all at one time; =**zeit**[˙ ¯], always, ever.

Figure 6

A relevant example is shown in Figure 6, listing the compound series based on the first element <all-> and starting with the entry <all||gebreichlich>.

A general occurrence may look like: lemma-1 in bold; constituent parts separated by double pipes <||> with no punctuation or space; accent rules included in brackets, followed by English translation (if more than one, they are separated by commas) and ended by semicolon. (Unspecified number of terms of the same type.)

Then lemma-2, in bold; the second part of the compound word only, preceded by <=> (small and oblique), then accent rules included in brackets, and English translation follows. Furthermore, for other entries whose word-formation is relevant, we are often given diminutives, comparison entries (<Cf.>), etymology or derivation preceded by the abbreviated source language <dG, E, F, G/HG, I, L, MHG, P, PG>[26] and the trace-back of the two words composing the sequence, separated by the symbol <+>. See for instance <allesmenanner>, in Figure 7.

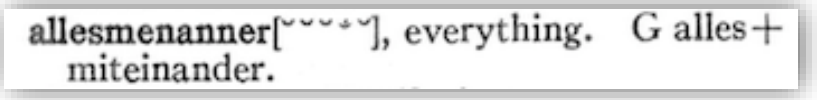

Figure 7

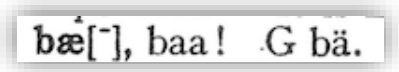

Figure 8

Nonetheless, the POS can be absent, see <bae>, in Figure 8, where the lexical category is omitted; there is no 'minimal standard' for the entries. Sometimes the entire line is limited to the lemma and its translation.

7. OCR scan phase

The overall semi-automatic process of text acquisition can be summarized by the following flow chart:

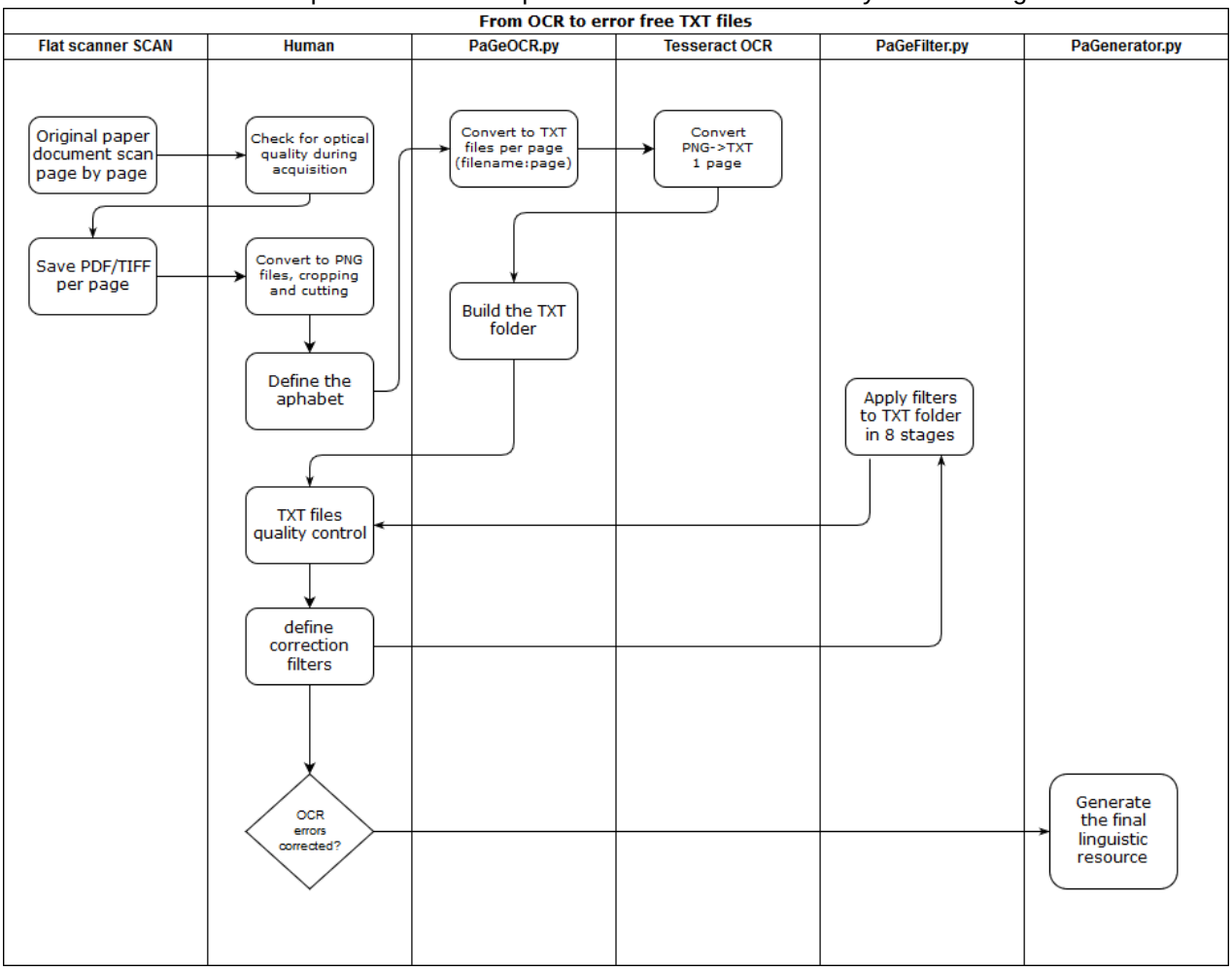

Figure 9

The overall technical process aims at minimizing the transcription errors while extracting text in the language of interest from books and printed documents, and as we have seen, Python programs have been developed to increase the accuracy of the semi-automated transcription, ideally aiming towards a 100% rate of

[26] Dialectic German, English, French, German/High German, Indian, Latin, Middle High German, Palatine dialect, Pennsylvania-German (Lambert 1924, XXIX).

correction accuracy in the final draft. Human intervention and the full integration of linguistics-based and IT-based knowledge were crucial for developing the loop, and in order to discover context dependent filters and design how to conveniently sequence them. The multi-stage process that is being tested in the project can be elaborated so to fix the OCR errors deriving from both the OCR reader and the features of the specific printed sources.

The underlying hypothesis is that those filters can incrementally be improved, with no conflicts, and could be applied to other printed sources displaying some comparable graphic appearance. For this reason, the filters have been divided into 8 banks or stages, each one dedicated to a level of abstraction of correction, from raw character sequence to n-gram context.[27] Therefore, the filter banks constitute *de facto* a knowledge base deriving from a rule-based contextual correction system.

See the flow chart, in Figure 9: the source paper document is scanned using a conventional high-resolution flatbed scanner at 120DPI, gray level acquisition, one page at a time. A human operator continuously checks the ideal optical conditions to obtain the best result, like black/white contrast and the absence of dust and particles. The scanner produces TIFF or PDF binary image files. The operator converts them into PNG files, conveniently cropping and cutting them, in order to isolate the text image to a single column of text, and then saves the files with names that index the page number. As the research team defines the recognition alphabet, the PNG dataset is fed to the PaGeOCR.py Python program, which converts it to TXT files using the Py-tesseract library, built upon the open source Tesseract software tool. So, an intermediate TXT folder is obtained, with file names indexing the page number as it is in the source. The dataset is, at this stage, reviewed and carefully scanned by the human operator, who identifies the errors and their occurrence frequency. It is then possible to divide them into classes, considering the mistake nature, features, and similarities and/or the similarity of the processes needed in order to solve them. This preliminary analysis provides a first definition of filters that shall be used in post-processing the OCR output. The entire set can be later refined, through a continuous observation of the out (result) after their application which is granted by the PaGeFilter.py Python program. When the accuracy is considered acceptable, the filtered TXT dataset is fed to the PaGenerator.py Python program that produces the linguistic resource in its final form, that is, in TEI (or spreadsheet formats). In this phase, the final program also takes care of sporadic spelling errors, which might still be present in the German and English segments, through the use of:

1. open source spell correctors, like the ones available from the LibreOffice project (which can be imported into Python programs),
2. Levenshtein function, which helps in fixing a spelling error when no more that 3 characters are missing or unexpected.

8. Classification of OCR errors

In a first phase, we led a preliminary recognition in order to evaluate the errors generated by the OCR scan, so as to divide them into general categories based on their nature (spelling errors, segmentation errors, specifically linguistic *impasses*); at a later stage only, it has been possible to identify specific areas of intervention and work on the abstraction of correction methodologies.

a) Character recognition errors

- Non-word detection and misrecognition of characters

When the OCR system fails in recognizing a character, an OCR error is produced, commonly causing a spelling mistake in the output text. The physical characteristics of the source (dirt on the page, flecks, stains and font-variations) preclude an accurate recognition of characters, which induce incorrect recognitions of words (e.g. <souiid> instead of <sound> or <&-Bi1rd#!> instead of <Bird>, <Z> instead of <2>). Other similar 'errors' can be Saxon genitive detection fail, words interrupted by a space, numbers taken for letters (<6.10> instead of <6.to>), <!> or <?> instead of some letter, space instead of <l>, and many more.

- Case Sensitivity

[27] An n-gram model is a type of probabilistic language model used to predict the next item in a sequence (of words, characters, etc). For language identification, sequences of *n* characters/graphemes (e.g., letters of the alphabet) are modeled (two letters or types: bigram, three letters or types: trigram) in order to find (statistically or linguistically) likely candidates for the correct spelling of a misspelled word.

OCR scanning often has difficulty in detecting uppercase and lowercase characters, and records incorrect entries such as, for example, <BrItaIn> or <BRITAIN>; we often faced the confusion of capital <i> in place of <l>, and many similar others.

- Lack of diacritical marks in the output-font

In the specific case study under review, we are dealing with a source (Lambert 1924) that makes use of many accents and diacritical signs. The output font does not include many of those, or finds them very difficult to recognize. A case in point is that of the tilde character, ~ [~], which must be understood as a suggestion on the correct (nasalized) pronunciation of the lemma entry, and which required much study and effort to be effectively tackled.

The entry <â ~ - > displayed in Figure 10, is a quite good example of it. The symbol <â> had to be introduced in the writing after the OCR scan phase. In this case it is followed by tilde [~] and minus <-> (signaling that the entry is a prefix); it is quite logic from a linguistic point of view, but at the same time very hard to detect. A closely related issue is the presence of the brackets, meant to contain the accents pattern. At this stage of research, the presence of brackets to be detected, analyzed, and corrected, raise issues that surpass that benefits of the information provided, so we have decided to get rid of them and do not include them in the thesaurus we wish to create.

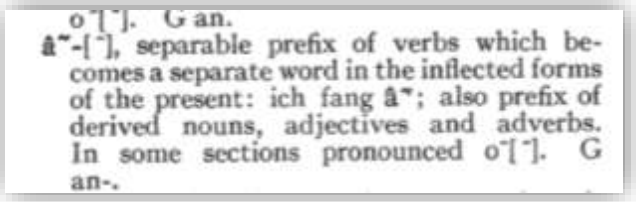
o ˜[ˉ]. G an.
â˜-[ˉ], separable prefix of verbs which becomes a separate word in the inflected forms of the present: ich fang â˜; also prefix of derived nouns, adjectives and adverbs. In some sections pronounced o ˜[ˉ]. G an-.

Figure 10

b) Segmentation errors

- Error in punctuation detection

Different line, word or character spacing leads to misrecognitions of white spaces, causing segmentation errors (e.g. <thisis> instead of <this is> or <depa rtmen t> instead of <department>). Mismatches of punctuation characters may also occur. Since we have generally used them to detect and isolate POS columns, errors of this kind turned out to be particularly insidious, because they could not but result in subsequent new errors in our classification.

- Tokenization errors and / or repetitions of meanings on the same entry

It is quite evident, see Figure 11, that a string like <aa/geh, v.; aagange, pp. 1.begin; 2.concern;> can be quite difficult to tokenize, to chop up into pieces – a *token* is an instance of a sequence of characters that can be grouped together as a useful semantic unit for processing – because it does not have any space between POS, point and semicolon. The same thing happens for numbers and definitions, i.e. <3.rave;>. The misrecognition of the numbers marking each new definition leads to the misunderstanding of the next "meaning" and to a redundancy – if not even a replication – of the entire line or following lines.

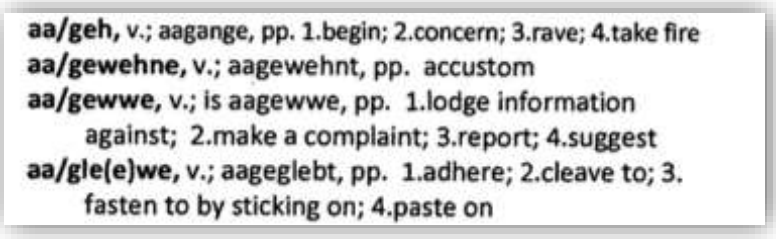
aa/geh, v.; aagange, pp. 1.begin; 2.concern; 3.rave; 4.take fire
aa/gewehne, v.; aagewehnt, pp. accustom
aa/gewwe, v.; is aagewwe, pp. 1.lodge information against; 2.make a complaint; 3.report; 4.suggest
aa/gle(e)we, v.; aageglebt, pp. 1.adhere; 2.cleave to; 3. fasten to by sticking on; 4.paste on

Figure 11

- Hyphenation errors

Tokens that are too long are split into line breaks. This case increases the number of segmentation errors exponentially. A very basic error correction algorithm is applied to reconstruct the text line integrity that can be corrupted by the OCR. It is based on a formal grammar rule parser that expects a typical text line sequence delimited by punctuation, when available, or by other contextual elements, like the current dictionary letter and textual context that can be recognized as beginning or end of the line.

c) Linguistic impasses

- Part-of-speech detection.

Often, as shown in Figure 12, the sources present definitions such as the following: <abard JMH, particular(ly), special(ly).> Besides the cold fact that Lambert does not specify the POS, this abundance of

parentheses happens because in this case Deitsch makes no difference between adjective and adverb. Of course, trying to build a bilingual tool, we need to consider that English does have a POS differentiation. Therefore, it is necessary to structure a system that is able to consider both options, and, finally, generate, in the inverse process, an 'asymmetric' link to the lemma in object.

- Regional variants, spelling and diacritical variants

There are sources that attest different spellings of terms already recorded and transcribed, and often also two or more orthographical variants within the same entry in the context of one single source (see again Figure 11 for <abbaddich> and its several graphies). It is necessary to continue the research in this sense, in order to ensure the greatest possible completeness to the final output, where all regional alternatives, spelling occurrences and diacritical variants will be included and linked to each other.

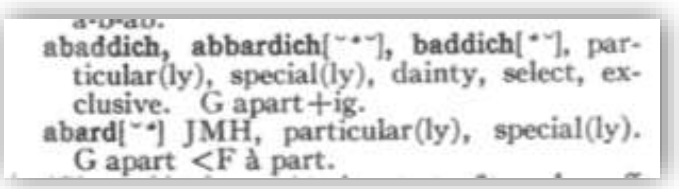
abaddich, abbardich[˘•˘], baddich[•˘], particular(ly), special(ly), dainty, select, exclusive. G apart+ig.
abard[˘•] JMH, particular(ly), special(ly). G apart <F à part.

Figure 12

9. Correction techniques

In order to provide for the correction of the detected errors, based on both their typology and the ratio of intervention needed for their resolution, we applied tools whose function is well known in literature, we implemented consolidated strategies, and studied new (above all corpus-focused) techniques.

a) Spell Corrector

The translations from PG to English present in the source dictionaries were processed by the software through the use of a spell checker[28] – acknowledged and widely validated – arriving at a resolution rate of 85% of the transcription errors related to those parts of text bearing the English translation of the entries. (Net of a percentage subjected to some new error, however – because the 'types' that are not recognized by the OCR can sometimes form new words, which are wrong within the context, but are correctly formed [e.g. 'mad' rather than 'sad'] and thus the spelling checker cannot recognize them as errors.)

b) Base Corrector

Part of the research involved conducting a percentage study on recurrent, and therefore easily avoidable, errors. It was possible to intervene through an ad-hoc structuring of a 'base corrector' able to filter the data sets from unnecessary punctuation and other redundant signs. The correction rules have been formulated on the base of both graphic and editorial characteristics of each source, and also depending on the configuration, organization and distribution of the single items.

c) Levenshtein distance or Minimum Edit Distance

Levenshtein distance or Minimum Edit Distance consists of identifying the minimum number of operations needed to convert a string into another string through insertion, deletion or replacement procedures. This technique was implemented in the Python post-processing OCR code in order to identify candidates with minimal modification distance, once the misspelled words had been identified.

d) Tree structured cycles

Most errors did not allow us to perform a simple replacement of a 'type' with another, in order to finalize the correction of the string. Each error case can generate incorrect characters or incorrect sequences of identical nature in different scenarios and contexts. Using criteria of subsequent approximations, we structured filters consisting of multiple selection levels with binary exclusion, leading, per stages, to a gradual correction of sometimes very complex errors. Here is a simple example: in place of <â / â ~> we found all of the following slightly different strings:

[28] PyEnchant, by Ryan Kelly: https://github.com/rfk/pyenchant. Python Library using Enchant by Dom Lachowicz: https://www.abisource.com/projects/enchant/.

fi’	é*	5"	é“	5 "
e’	fi"	é’	fl’	fif
fi	a’	fi”	a’	5’
a“	è	é.’	5.	a”
é”	fi’	a’	5.”	§“
è	a	ii"	55"	2’1”

Narrowing down to the binary choice <â / â ~> implied working on strings bearing a variable number of characters, 1 to 5, often recurring also elsewhere, for instance <fl’> in place of <fl> in the English word <fly>; or <5.> in its correct form and use, as it would be for introducing the fifth definition of a word (and in this case with no need to be corrected for having no relation with neither <â / â ~>.)

e) Probabilistic techniques

Confusion probabilities: there were a large number of other cases in which a sequence had been frequently mistaken in a ‘systematic’ way, generating repetitive errors. (E.g.: <p1> instead of <pl> for ‘plural’ and many others.) In those cases, a modification was opted for the whole string.

e.g.: <11 : n> <111 : m><1’, 1’: f><0f : of><D1m : dim><1(ly) : l(ly><z1e : zle><k1e : kle><1n : in><Ad] : Adj><t1g : tig><1t : it><slch : sich><F rom : From><9p : pp><(p1) : (pl)><tln : tin><N oun : Noun>.

However, most of the work was carried out relying on linguistic rules: *Transition probabilities* represent the probabilities that a given letter (or sequence of letters) is followed or not followed by another specific character. Transition probabilities strictly depend on the language and in some cases on the spelling conventions adopted by the author. For example, Deitsch words do not end with the following sequences: lr, tr, kr, fr. So these sequences are more likely to be misrecognitions than other strings, in this case: <lr : l [>, <tr : t [>, <kr : k [>, <fr : f [>.

Another very simple example is the impossible letter sequence <sclll> found in the OCR output. This sequence was found in 8 Deitsch headwords like <scllpengler> (in one case the sequence was <scllll>: <wesclllltlich>) and could be easily transformed automatically in the correct sequence <sch>.

A more intriguing issue is the erroneous detection of <ü> in Deitsch lemmata, a letter that normally is not used in Deitsch since the corresponding German vowel does not exist (though it occurs in the German cognates). At a careful analysis of the instances, there appear to be different underlying situations producing the same output error, and therefore different approaches are needed, and, most importantly, a logically ordered sequence of the concurrent correction steps.

1. <u>In word final position</u>

- In cases like <degleicherü> or <verunglickerüw> in place of <degleiche> and <verunglicke>. The word final sequences <rü>, <rüw> (and similar ones) are clearly due to a misrecognition of the square brackets <[]> located after the lemma. These strings can be corrected separately as a whole, because, of course, the misrecognition of the square brackets does not lead to the sole false strings <rü>, <rüw> but also to many others, and they all will be fixed together, in a specific stage of correction.
- some occurrences display letter-sequences like <besarrickü> and <schâmgrauü>, where <ü> stands for <t>.
- In all the other instances, <ü> stands for <ff>: <druü> is <druff>, <schiü // bruch> is <schiff // bruch> and so on.
  Being <ü> for <t> less frequent, it is necessary to correct these two particular words before implementing the automatic correction rule which transforms the final <ü> in final <ff>.

2. <u>In word initial or internal position</u>

- There are single errors which appear in just one item: <würfe> instead of <wærfe> and <drücke> instead of <drucke>. In these cases, it is not possible to generate a correction rule but the single item has to be corrected.

- It is observed that in all cases where <ü> is followed by <a> or <u> the correct sequences are <fla> and <flu>: <abüaddre> should be <abfladdre>; < abüuche> stands for <abfluche>. This allows the formulation of the following correction rule: <üa> : <fla> and <üu> : <flu>.
- All the remaining instances of erroneous <ü> need to be transformed in <fi>: <ausüsche> must be <ausfische>, <üngerling> should be <fingerling>, <lâüch> is <lâfich>, and so on.
- With one exception: <wiüelt> must be <wiffelt>. All exception detected, at any case, need to be corrected before generalizing the above-mentioned rule in order to avoid the production of new errors.

f) Verification of corrections

A tool was developed, proficient in extracting a lemma at a time according to random criteria (Random Picker). It provides a line with the physical location of the term, followed by the 'raw' lemma, as detected and recorded by the OCR, followed by all the correction levels. See Figure 13.

```
143a_143b.txt 122
DA: schtengel H bauer[""'"], m, poor farmer; =g1äs [W'], n, Wineglass.
0 : schtengel // *bauer[""'"], m, poor farmer; =g1äs [W'], n, Wineglass.          H : //
1 : schtengel // *bauer, m, poor farmer; =g1äs , n, Wineglass.                    [-]:
2 : schtengel // *bauer, m, poor farmer; =g1äs , n, Wineglass.
3 : schtengel // *bauer, m, poor farmer; =g1äs , n, Wineglass.
4 : schtengel // *bauer, m, poor farmer; =g1äs , n, Wineglass.
5 : schtengel // *bauer, m, poor farmer; =g1äs , n, Wineglass.
6 : schtengel // *bauer, m, poor farmer; =g1âs , n, Wineglass.                    ä:â
7 : schtengel // *bauer, m, poor farmer; =g1âs , n, Wineglass.
8 : schtengel // *bauer, m, poor farmer; =glâs , n, Wineglass.                    g1a:gla
```

Figure 13

Each line corresponds to an output string indicating which correction rule has been applied at each step, and the corresponding result in the modification process of the primary entry. The Random Picker is very useful in processing large random quantities of lemmas and verifying the adequacy and precision of the multi-stage filters set.

10. Achieved results

The study of PG linguistics allowed the codification of rules that gradually perfected the electronic output produced thanks to the use of Python. The software so developed has therefore produced files in Excel format that reproduce in digital form the analyzed dictionaries, including all the signs which are necessary to the univocal identification of the lemma, and its correct pronunciation. A grammatical and syntactic classification – internationally recognized in German computational linguistics (POS-tagging according to the STTS system proposed in Schiller, Teufel and Stöckert 1999) – has been introduced for each entry or compound, and each entry corresponds to a translation in English, and in some cases also to an etymology, an example or micro-context, a diminutive and/or a German equivalent.

11. Research continuation and integration

The results so far obtained suggest and configure the possibility – once all the available sources shall be analyzed and digitized – of structuring an accessible online platform, which will serve as a translation tool and research device for both English and PG-based studies, being at the same time an archive of the various spellings and / or dialectal forms attested in PG.

The resource is also a place of conservation and dissemination of a minority cultural heritage of the United States, thus contributing to a better understanding of the American multilingual landscape and to the dynamics of contact and exchange.

The theoretical value of our research is determined by its strictly applied linguistic competence, integrated by the vast possibilities of the Python programming language. The study conducted on these assumptions allowed, through "The PAGe" project, the subsequent structuring of an 'intelligent' automatic correction

model, which constitutes a method that is widely reproducible and applicable to other corpora, also in other languages.

"The PAGe," therefore, also aims at being a flexible protocol model for the approach to the correction and systematization of the digitization process of sources, by the means of textual analysis. The single phases and the processes developed for the correction/classification operations can be replicated, adapting from time to time the conceptual model to the specific corpus.

Acknowledgements

We would like to express our gratitude to Giovanni De Gasperis and Gianna Fusco for their fundamental contribution to the research project and to the draft of this paper.

References

Beam, C. Richard and Joshua R. Brown / Jennifer L. Trout. *The Comprehensive Pennsylvania German Dictionary*. 12 voll. A. Morgantown: Masthof, 2004-2011.

Bloomberg, Dan S. "Determining the Resolution of Scanned Document Images." *Document Recognition and Retrieval* VI.3651 (1999): 10-22.

Buffington, Albert F., and Preston Albert Barba. *A Pennsylvanian German Grammar*. Allentown: Schlechter's, 1954.

Fishman, Joshua. "Language Maintenance and Language Shift as Fields of Inquiry: A Definition of the Field and Suggestions for Further Development." *Linguistics* 9 (1964): 32-70.

Frey, J. William. *A Simple Grammar of Pennsylvania Dutch*. Lancaster: Brookshire, 2009.

Fuller, Janet M. "When Cultural Maintenance Means Linguistic Convergence: Pennsylvania German Evidence for the Matrix Language Turnover Hypothesis." *Language in Society* 25.4 (1996): 493-514.

Haag, Earl C. *A Pennsylvania German Reader and Grammar*. University Park and London: Pennsylvania State University, 1982.

Hans-Bianchi, Barbara. "Pennsylvaniadeutsch: Wege der Verschriftung einer Minderheitensprache." *Baig* VII (2014): 113-131. http://www.associazioneitalianagermanistica.it/images/bollettini/9_Hans-Bianchi_113-131_DEF.pdf. Last visited 14/12/18.

---. "Kodifizierung als Überlebensstrategie? Orthographische Kodifizierungsversuche in Pennsylvania Deitsch." *Die Kodifizierung der Sprache. Strukturen, Funktionen, Konsequenzen* (= WespA, vol. 17). Eds. Wolf Peter Klein and Sven Staffeldt. 2016. 42-69. https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/13808/file/WespA17_Kodex_Klein_Staffeldt.pdf

Horne, Abraham R. *Horne's Pennsylvania German Manual*. Allentown: Horne, 1875.

Rauch, Edward H. *Pennsylvania Dutch Hand-book / Pennsylvania Deitsh Hond-Booch*. Mauch Chunk: Rauch, 1879.

Keiser, Steven Hartman. *Pennsylvania German in the American Midwest*. Durham: Duke University Press, 2012.

Klein, Wolf Peter. "Gibt es einen Kodex für die Grammatik des Neuhochdeutschen und, wenn ja, wie viele? Oder: Ein Plädoyer für Sprachkodexforschung." *Sprachverfall? Dynamik Wandel Variation*. Eds. Albrecht Plewnia and Andreas Witt. Berlin: Mouton de Gruyter, 2014. 219-242.

Lambert, Marcus B. *A Dictionary of the Non-English Words in the Pennsylvania-German Dialect*. Lancaster: Pennsylvania German Society, 1924.

Louden, Mark L. "Pennsylvania German in the Twenty-first Century." *Sprachinselwelten. Entwicklung und Beschreibung der deutschen Sprachinseln am Anfang des 21. Jahrhunderts*. Eds. Nina Berend and Elisabeth Knipf-Komlósi. Bern: Peter Lang, 2006. 89-107.

---. *Pennsylvania Dutch. The Story of an American Language*. Baltimore: Johns Hopkins University Press. 2016.

Schiller, Anne, Simone Teufel and Christine Stöckert. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. 1999. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf. Last visited 14/12/18.

Seifert, Lester W. J. “The Word Geography of Pennsylvania German: Extent and Causes.” *A Word Atlas of Pennsylvania German*. Eds. Mark L. Louden, Howard Martin and Joseph C. Salmons. University of Wisconsin-Madison: Max Kade Institute, 2001. 81-102.

Springmann, Uwe. “A High Accuracy OCR Method to Convert Early Printings into Digital Text.” 2015. http://cistern.cis.lmu.de/ocrocis/tutorial.pdf

Stine, Eugene S. *Pennsylvania German Dictionary: Pennsylvania German-English, English-Pennsylvania German*. Kutztown: Pennsylvania German Society, 1996.

Stolberg, Doris. *Changes between the Lines*. *Diachronic Contact Phenomena in Written Pennsylvania German*. Berlin: Mouton de Gruyter, 2015.

Van Pottelberge, Jeroen. *Der* am-*Progressiv. Struktur und Parallele Entwicklung in den Kontinental-Westgermanischen Sprachen*. Tübingen: Narr, 2004.

Villa, Daniel J. and Susana V. Rivera-Mills. “An Integrated Multi-generational Model for Language Maintenance and Shift. The Case of Spanish in the Southwest”. *Spanish in Context* 6:1 (2009): 26-42.

Werner, Michael. *Lexikalische Sprachkontaktphänomene in Schriftlichen Texten des Pennsylvaniadeutschen*. Ph.D. dissertation. University of Mannheim. 1996.

Zacharias, Peter. *Pennsylvania Dutch Dictionary*. (Online dictionary based on Lambert 1924). https://www.padutchdictionary.com. Last visited 14/12/18.