Mandenkan
Bulletin semestriel d'études linguistiques mandé

59 | 2018
Numéro 59

Texts for the corpus of Nko: collection, conversion, and open issues

Textes pour le corpus de n'ko: collection, conversion et problèmes pendants

ТЕКСТЫ ДЛЯ КОРПУСА НКО: СБОР, КОНВЕРТАЦИЯ И ОТКРЫТЫЕ ВОПРОСЫ

Andrij Rovenchak

Electronic version
URL: https://journals.openedition.org/mandenkan/1360
DOI: 10.4000/mandenkan.1360
ISSN: 2104-371X

Publisher
Llacan UMR 8135 CNRS/Inalco

ELECTRONIC REFERENCE

Andrij Rovenchak, **"Texts for the corpus of Nko: collection, conversion, and open issues"**, *Mandenkan* [Online], 59 | 2018, Online since 20 July 2018, connection on 08 July 2021. URL: http://journals.openedition.org/mandenkan/1360 ; DOI: https://doi.org/10.4000/mandenkan.1360

Texts for the corpus of Nko: collection, conversion, and open issues[1]

Textes pour le corpus de n'ko: collection, conversion et problèmes pendants

ТЕКСТЫ ДЛЯ КОРПУСА НКО: СБОР, КОНВЕРТАЦИЯ И ОТКРЫТЫЕ ВОПРОСЫ

Andrij Rovenchak

Acknowledgments. *I am grateful to the anonymous referees for their comments which induced me to extend the initially concise presentation of the material. I appreciate Christopher Green's and Dafydd Gibbon's help with language editing of the paper.*

Introduction

1 The Nko alphabet was created in 1949 by a Guinean scholar and enlightener Solomana Kantɛ (1922–1987) as a script for Manding languages. It is the most successful new indigenous African writing system currently in use (Oyler 2005; Vydrine 2001). The script is mainly utilized for the Maninka language of Guinea.

2 Preliminary steps in compiling a corpus of Maninka, as well as the description of the Maninka Reference Corpus (Corpus Maninka de Référence, CMR), where the vast majority (over 85 percent) of texts are written in Nko, have been described in a series of papers (Davydov 2010; Vydrin 2013; Vydrin 2014; Vydrin, Rovenchak & Maslinsky 2016).

3 The present report briefly summarizes the experience with the Nko corpus building and outlines prospects towards further development in the future.

Collection of texts

4 We started collecting texts for the corpus of Nko in 2009. A number of Nko texts in electronic form are available from web sites (mainly http://www.kanjamadi.com) and social networks. The latter comprise a modest fraction of all the texts in the corpus, but such texts (Twitter and Facebook posts) are rather specific and, consequently, they have been the focus of recent studies (see Mikros & Perifanos 2015 and references therein). Presently, more than 1,000 tweets in Nko can be collected. The authors tweeting mostly in Nko include @solofarabado (about 500 tweets), @kiniebaka (over 300 tweets), and @siarabohlansine (over 100 tweets). Other online media, for instance, the NkoAfrica blog (http://nkoafrica.over-blog.com) launched in 2017, represent a good source for potential corpus expansion. Several pages from the Wikipedia incubator in Nko (https://incubator.wikimedia.org/wiki/ Special:PrefixIndex/Wp/nqo/) might be also included. A large part of Nko texts in electronic form (in DOC and PDF formats) were obtained from the Nko Academy (*Ńkó Dúnbu,* http://nkodoumbou.com) via Ibrahima Sory 2 Condé. All the files were converted to a plain-text format in the standard UTF-8 encoding using software specially developed by the author.

5 As of April 14, 2018 (the day marking the 69th anniversary of Nko), the corpus size is 3,122,178 words (Vydrin et al. 2014). Over 20 percent of the corpus (ca. 650,000 running words) is covered by texts from periodicals,[2] in particular, *Dàlu Kéndɛ* and *Yèreya fɔ̀ɔ̀bɛ* (cf. Rovenchak 2015a). Religious texts (mostly Qur'an and Tafsīr) count over 500,000 words. The remaining part of the corpus is composed of educational literature, texts of popular nature, works of fiction, etc. The corpus can be freely accessed online at http://cormand.huma-num.fr/cormani/.

Text conversion

6 The conversion of texts from PDF requires several steps. First of all, the text is copied from PDF into a text file in a proper order, which is particularly important for multicolumn periodicals, and break points are inserted for subsequent splitting of longer texts into individual articles or sections. This last step justifies the manual processing of PDFs: note that article titles are often given as WordArt objects or similar images and not as ordinary text and thus require re-typing. At present, we have processed over 100 issues of periodicals which constitute about one third of the available files. The obtained text files are then converted into UTF-8 format. Various encoding schemes are attested when processing original PDFs; there are currently approximately ten approaches used. Occasionally, two different encodings appear in one PDF for different text blocks.

7 Some of the most typical encoding schemes are summarized below. For internal purposes, they are denoted by abbreviations (NN, NN1, NN3, etc.). The scheme NN3 is attested, in particular, for *Dàlu Kéndɛ* from №15 (15 August 2011) to №25 (31 October 2011). Figure 1 shows text copied from PDF (left image) as compared to the converted one (right image).

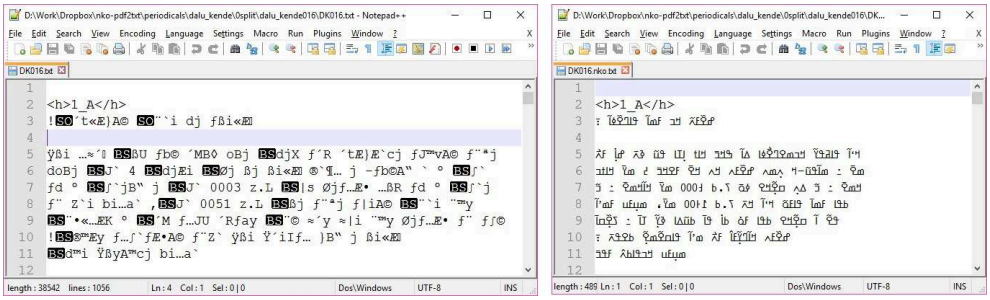


Figure 1: Screenshots of texts from *Dàlu Kéndɛ* from №16 (22 August 2011).

Note technical headers within tags <h>…</h> used to make automatic splitting of entire issues into separate articles. The left panel is the original copied text while the right is the text in a Unicode font.

8 A screenshot of a fragment of the Perl script used to make the conversion to Unicode is given in Fig. 2.

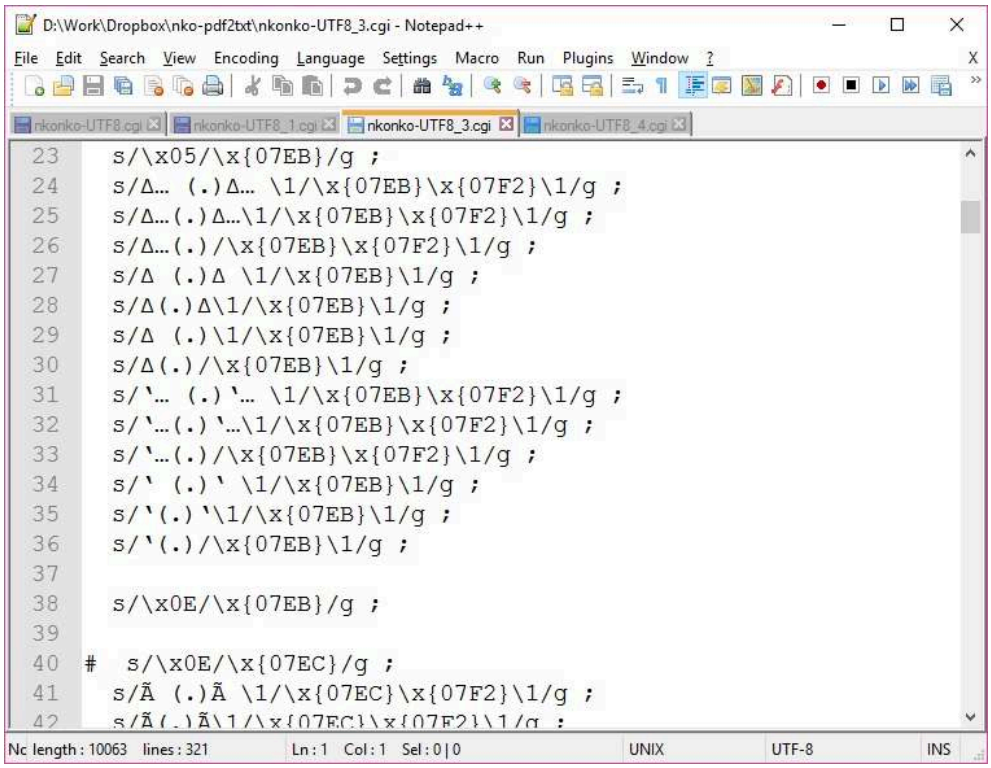


Figure 2: A fragment of the Perl script for the conversion to the Unicode format from the NN3 encoding.

9 Another encoding scheme (NN) is demonstrated in Fig. 3. Note that system fonts do not contain codepoints of the text copied from PDF, so we can see just empty boxes. A conversion script fragment is shown in Fig. 4. This scheme is attested in particular for *Dàlu Kéndɛ* from №96 (20 May 2013) to №99 (10 June 2013).

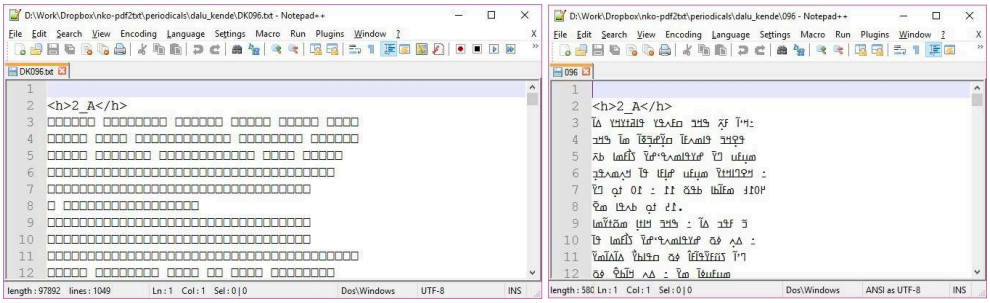


Figure 3: Screenshots of texts from *Dàlu Kéndɛ* from №96 (20 May 2013).

The left panel is the original copied text while the right is the text in a Unicode font.

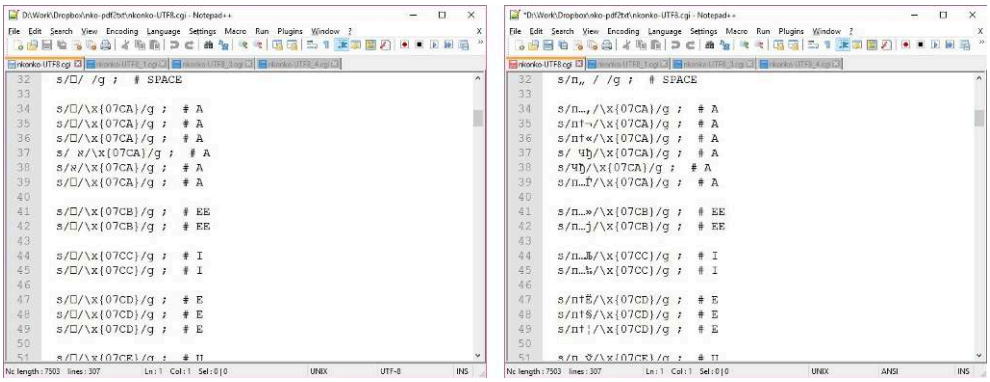
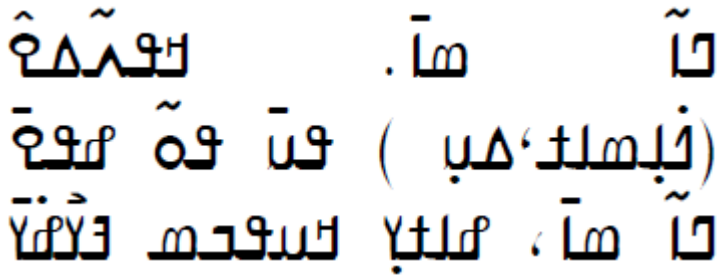


Figure 4: A fragment of the Perl script for the conversion to Unicode format from NN encoding.

The left panel corresponds to the UTF-8 mode while the right is the ANSI mode of text representation. Note in the top row an unusual encoding of the space character.

10 While processing the files, several typical misprints were discovered. They often do not cause any problems in reading and understanding the printed version, but they hinder straightforward automatic processing of texts.

11 For instance, the *dágbasinna* symbol < ◌ٜ > (U+07D1) cancelling the *gba◌rali* (contraction) orthographic rule is often replaced by the low tone apostrophe < ‘ > (U+07F5), as in the following example (the rightmost word in the middle line, *ʒándarmun*):



[...]	na◌	dá.	kɛ̀lɛ-mɔ́ɔ̀	(ʒándarmun)	lú	lɛ̀	fɔ́lɔ
[...]	come	PFV.INTR	war-human-ART	(FRGN)	PL	FOC	first

na◌̀	dá.	fárin	kúlod	pi̋ivi	[...]		
come	PFV.INTR	vigorous	Claude	Pivi	[...][3]		

12 Note that the corpus is glossed using the Daba software (Maslinsky 2014) based on the Malidaba dictionary (Vydrin, Rovenchak & Maslinsky 2016). The preparation of a cleaned, disambiguated, and thoroughly checked Nko corpus is a subsequent stage, which would require resolving some of the issues mentioned below.

13 Additional spacing is not rare in combination with tonal apostrophes and tonal marks. Moreover, in many cases, the placement of tonal marks (before or after a vowel, as they are stored in the PDFs) varies in the same file. A fully automatic standardization of such situations is not possible due to appearance of foreign sounds that use diacritical marks, cf. <ߊ߫ߓߘߎߟߊ߯ߤ߭ߌ> *ábdulaaȟi* ‘Abdulaaȟi’ or <ߛߊ߫ߙߞߏߖ߭ߌ> *sárkozi* ‘Sarkozy’. In some cases, it is not immediately obvious whether the mark belongs to the consonant or to the neighboring vowel, cf. <ߌ߫ߙߊߞߌ> *íraki* ‘Iraq’, where the overline denotes the high tone of the final syllable, versus <ߌ߫ߙߊߞ߫ߌ> *íraqí*` ‘Iraq’, where the overline changes < ߞ > /k/ to /q/. These problems are often observed in periodicals, where foreign names occur rather frequently. Manual correction seems to be the best option so far. A Nko spellchecker, which would facilitate handling of the abovementioned misprints, is under development by Jean Jacques Méric (cf. Méric 2014 describing the Bamana spellchecker).

14 Another curious and unexpected typographic solution is the “nasalized” Arabic comma (dot below < ،߲ >, U+060C followed by U+07F2), occurring instead of the Arabic semicolon < ؛ > (U+061B). Such situations are easily handled automatically.

15 In February 2016, an online converter between various Nko encodings, previously determined while processing DOC files for the corpus, was launched at http://nkoconvert.ho.ua. Initially, it was intended as a tool for conversion between older (pre-Unicode) fonts and Unicode format. Two different mappings of Nko letters on Arabic letters were used in those fonts, one of which is represented by the following typefaces: *BOURAMA-KANTE, Nko Manding1 Cote D'Ivoiredf1, N'ko2 Manden-1, Solomana Kante*, while the second is found in *A Manding BATEKA, A Manding IT BAMA, Fofona, Karifala Berete, Manding N'ko Sigui, Nko Africa, Nko Kouroukan Fouwa*, and some others (Vydrin, Rovenchak & Maslinsky 2016). Mapping on another right-to-left script, Hebrew, is used in the *NkoHeb* typeface, which is not frequently attested in available materials.

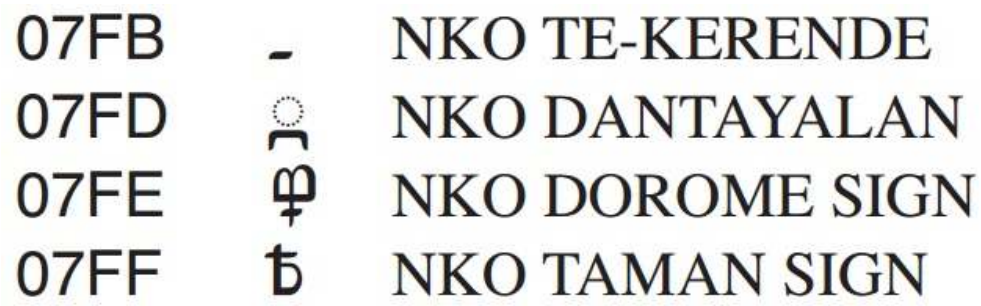
16 Other functions of this converter include the transformation between the Nko script and the Roman-based orthography (with full notation of tones) and the transformation from the Roman based orthography (without tones) to Nko. There are certain limitations in the latter conversion. Since the original Roman-based texts have no tonal notation, the Nko version of the respective text cannot be tonalized. Short vowels have no tonal marks, i.e., they are represented as if they would carry a high tone. Taking into account frequency data about tonal patterns in Nko texts (Rovenchak 2011; Rovenchak 2015b), one can estimate that about 70% of vowels would be represented correctly. A similar decision was made to mark a long vowel by the diacritic mark corresponding to “long vowel + high tone”. The only exception made is for the long vowel /ɔɔ/, which appears in Maninka much more often with low tone. Since there is no tonal distinction,

the *gbàrali* (orthographic vowel contraction) rule is not applied in the Nko version of texts converted from the Roman-based orthography. In the future, more advanced techniques might be applied similar to those used by Liu & Nouvel (2017) for Bamana texts.

Open issues

17 A number of open issues remain in order to facilitate further acquisition of texts for the Nko corpus. They are of different natures and importance, and only a couple of them are outlined below.

18 Some minor issues include the possibility of encoding additional characters in the Nko Unicode block (Everson 2015):



19 Of the four proposed symbols, the *tɛ́-kɛrɛndɛ* mark is most frequently needed. Its role is similar to a hyphen in some compound words; presently, several different approaches are used to represent this symbol. The combining *dántayalan* mark denoting abbreviations of units of measure also occurs in the corpus several times. The remaining two signs, which are currency symbols, have not been attested in the corpus so far. The Unicode Technical Committee suggested that the *tɛ́-kɛrɛndɛ* mark be represented by a hyphen (Anderson et al. 2016). A further, detailed study of Nko texts is required to determine whether there is a crucial difference between *tɛ́-kɛrɛndɛ* and hyphen. So far, this mark can be represented as an underscore < _ > (U+005F). The other three characters (*dántalayan*, *dɔ́rɔmɛ*, and *táman*) are expected to appear in the Unicode Version 11.0 scheduled[4] for release in June 2018.

20 The currently available, though incomplete conversion from the Roman-based to the Nko orthography requires further tuning to reflect tonal marking in Nko texts. Such an algorithm will heavily rely on a dictionary, which is under development (Vydrin, Rovenchak & Maslinsky 2016).

21 In PDF documents, especially those produced using modern Unicode fonts, problems with text block placement occur frequently. Namely, even a single text line can appear as a set of non-consecutive blocks. While it might be unexpected, the problem could originate in PDF algorithms for representing right-to-left scripts. Therefore, to facilitate further collection of texts, it would be practical if sources for published texts (in DOC, DOCX, ODT, or TXT formats) were made available by authors or publishers, rather than solely the respective final PDFs. Obviously, no text will be put online in its entirety but will be used solely for processing towards the inclusion into the corpus.

22 In the future, it would be useful to implement optical character recognition (OCR) for Nko as such software is not currently available. One possibility is to train an open-source OCR engine known as Tesseract (https://github.com/tesseract-ocr) for the

Nko script. OCR software would allow processing of both printed materials not available electronically and of PDF documents lacking text layers for current conversion algorithms.

Concluding remarks

23 I expect the presented information to generate feedback from the Nko community and hope that it plants the seeds for further collaborations in collecting texts for the Nko corpus as well as in improving existing tools for processing Nko texts. Separate attention should be paid to texts available in several languages as being a source for multilingual parallel corpora; these will be invaluable in the development of automatic translation services in the future.

BIBLIOGRAPHY

References

Anderson, Deborah, Ken Whistler, Rick McGowan, Roozbeh Pournader, Andrew Glass & Laurentiu Iancu. 2016. *Recommendations to UTC #146 January 2016 on Script Proposals. L2/16 037*. https://www.unicode.org/L2/L2016/16037-script-rec.pdf.

Davydov, Artem. 2010. Towards the Manding corpus: Texts selection, principles and metatext markup. In Guy de Pauw, Handré Groenewald & Gilles Maurice de Schryver (eds.), *Proceedings of the Second Workshop on African Language Technology AfLaT, May 18, 2010*, 59–62. Valletta, Malta. http://tshwanedje.com/publications/AfLaT2010.pdf.

Everson, Michael. 2015. *Proposal to encode four N'Ko characters in the BMP of the UCS. L2/15 338*. http://www.unicode.org/L2/L2015/15338-n4706-nko-additions.pdf.

Liu, Yu Cheng & Damien Nouvel. 2017. A Bambara tonalization system for word sense disambiguation using differential coding, segmentation and edit operation filtering. *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017), Nov 2017*, 694–703. Taipei, Taiwan.

Maslinsky, Kirill. 2014. Daba: a model and tools for Manding corpora. In Mathieu Mangeot & Fatiha Sadat (eds.), *Actes de l'atelier sur le traitement automatique des langues africaines TALAf 2014 (Actes des Ateliers TALN 2014. Éd. par Brigitte Bigi)*. http://www.aclweb.org/anthology/W14-6502.

Méric, Jean Jacques. 2014. Un vérificateur orthographique pour la langue bambara. In Mathieu Mangeot & Fatiha Sadat (eds.), *Actes de l'atelier sur le traitement automatique des langues africaines TALAf 2014 (Actes des Ateliers TALN 2014. Éd. par Brigitte Bigi)*. http://www.aclweb.org/anthology/W14-6505.

Mikros, George K. & Kostas Perifanos. 2015. Gender identification in Modern Greek tweets. In Arjuna Tuzzi, Martina Benešová & Ján Mačutek (eds.), *Recent Contributions to Quantitative Linguistics*, 75–88. Berlin–Boston: Mouton de Gruyter.

Oyler, Dianne White. 2005. *The history of the N'Ko alphabet and its role in Mande transnational identity: Words as weapons*. Cherry Hill, NJ: Africana Homestead Legacy Press.

Rovenchak, Andrij. 2011. Phoneme distribution, syllabic structure, and tonal patterns in Nko texts. *Mandenkan* 47. 77–96.

Rovenchak, Andrij. 2015a. Quantitative studies in the corpus of Nko periodicals. In Arjuna Tuzzi, Martina Benešová & Ján Mačutek (eds.), *Recent Contributions to Quantitative Linguistics*, 125–138. Berlin–Boston: Mouton de Gruyter.

Rovenchak, Andrij. 2015b. Quantitative studies in the corpus of Nko periodicals. In Arjuna Tuzzi, Martina Benešová & Ján Mačutek (eds.), *Recent Contributions to Quantitative Linguistics*, 125–138. Berlin–Boston: Mouton de Gruyter.

Vydrin, Valentin. 2013. Bamana Reference Corpus (BRC). *Procedia – Social and Behavioral Sciences* 95. 75–80.

Vydrin, Valentin. 2014. Projet des corpus écrits des langues manding : le bambara, le maninka. In Mathieu Mangeot & Fatiha Sadat (eds.), *Actes de l'atelier sur le traitement automatique des langues africaines TALAf 2014 (Actes des Ateliers TALN 2014. Éd. par Brigitte Bigi).* http://www.aclweb.org/anthology/W14-6501.

Vydrin, Valentin, Kirill Maslinsky, Aandrij Rovenchak & Ibrahima Sory 2 Condé. 2014. Corpus maninka de référence. Corpus annoté des textes écrits en maninka, surtout en écriture Nko. http://cormand.huma-num.fr/cormani/.

Vydrin, Valentin. Andrij Rovenchak & Kirill Maslinsky. 2016. Maninka Reference Corpus: A Presentation. *Actes de la conférence conjointe JEP TALN RECITAL 2016, volume 11: TALAF*, 87–94. Paris: AFCP et ATALA.

Vydrine, Valentin. 2001. Soulemane Kantè, un philosophe innovateur traditionaliste maninka, vu à travers ses écrits en Nko. *Mande Studies* 3. 99–131.

NOTES

1. This material was intended for ALDP 2017 *8th African Languages in the Disciplines and Professions' Conference* (Conakry, Guinea, 21-23 April 2017).
2. Electronic versions are available from the Nko Electronic Library (http://cormand.huma-num.fr/maninkabiblio/index.jsp) or Kanjamadi (http://kanjamadi.com/foobe.htm).
3. I am grateful to Valentin Vydrin for assistance with this example glossing. PFV.INTR corresponds to the intransitive perfect mark, ART denotes the tonal article, FRGN is the gloss for a loanword, PL is the plural mark, and FOC is the contrastive focus mark.
4. In the Unicode Version 11.00 released on 05 June 2018 the following codepoints are assigned: U+07FD NKO DANTAYALAN; U+07FE NKO DOROME SIGN; U+07FF NKO TAMAN SIGN.

ABSTRACTS

This paper discusses the compilation of a Maninka corpus, where the majority of texts are written in the Nko alphabet. Prospects for further development of the Nko corpus are briefly outlined. With this information presented, the aim is to further extend collaborations in collecting texts for the Nko corpus and to improve existing tools for processing Nko texts.

L'article présente une expérience de compilation du corpus maninka, langue où la majorité des textes sont écrits en alphabet n'ko. Les perspectives de développement du corpus n'ko sont brièvement décrites. Les informations présentées visent à étendre davantage la collaboration dans la collection de textes pour le corpus n'ko et dans l'amélioration des outils existants pour le traitement des textes en n'ko.

В СТАТЬЕ ПРЕДСТАВЛЕН ОПЫТ СОСТАВЛЕНИЯ КОРПУСА ЯЗЫКА МАНИНКА, ГДЕ БОЛЬШИНСТВО ТЕКСТОВ НАПИСАНО НА АЛФАВИТЕ НКО. КРАТКО ИЗЛОЖЕНЫ ПЕРСПЕКТИВЫ РАЗВИТИЯ КОРПУСА НКО. ПРЕДСТАВЛЕННАЯ ИНФОРМАЦИЯ НАПРАВЛЕНА НА ДАЛЬНЕЙШЕЕ РАСШИРЕНИЕ СОТРУДНИЧЕСТВА В СБОРЕ ТЕКСТОВ ДЛЯ КОРПУСА НКО И СОВЕРШЕНСТВОВАНИЕ СУЩЕСТВУЮЩИХ ИНСТРУМЕНТОВ ОБРАБОТКИ ТЕКСТОВ НА НКО.

INDEX

motsclesru НКО, КОРПУС ТЕКСТОВ, МАНИНКА, КОДИРОВКА, ЮНИКОД
Keywords: Nko, text corpus, Maninka, encoding, Unicode
Mots-clés: N'ko, corpus de textes, maninka, codage, Unicode

AUTHOR

ANDRIJ ROVENCHAK

АНДРИЙ РОВЕНЧАК
Ivan Franko National University of Lviv, Ukraine
andrij.rovenchak@gmail.com