日エス辞典（宮本正男編）の電子テキストデータの修正作業について

２０１６．５．２９

不十分ですが最初に簡単に修正する項目を挙げます。詳しくは後半を見てください。

◎　ページの変わる所に、ページの左右を示す特別な行を入れていますが、その他は印刷された辞書の行と一致するように改行したテキストデータです。

「かな」文字の変わるところの空白には　◎◎◎◎◎　を４行入れ、各頁の左右がそれぞれ６４行になるようにしています。

◎　見出し部分を　★で囲み、行の終わりは　⏎、単語の終わりは　◆⏎　です。

ただし、説明が 1000 文字を超える場合は、見出しに番号を付け途中で分けているため、印刷の行とずれがあります。（後半に説明あり）

◎　単語などの区分がはっきりするように、空白は全角スペースにしています。

◎　一部のかっこなどの記号は、印刷物と変えていますので、特に注意してください。

見出しの漢字部分、専門分野、説明の大区分の数字、参照の矢印など

例：　★ 02G03 やま【山】★　〖１〗　〖地〗　monto.　〔大きい〕　montego.　〔小さ⏎

い〕　monteto.　※　～の多い　montoriĉa.　※　～に登る　grimpi　⏎

略　　（※マークは、プログラムで、例文・派生語ごとに改行するために追加）

diveno　ĝuste　trafis　en　ekzameno.　※　〖５〗　〔頂点〕　⇒　や⏎

まば.　◆⏎　　（補足説明の日本語は、同じ〔　〕　のまま）

◎　エス単語の出典を示す小さな添え字 I　K などは、*I　*K とし、語の前に付いているときは　*I*　*K*　とする。 * の場合はそのまま。

◎　動植物の学名は　〖　〗　に変更

★ 15G43 ヤマカガシ★　〖動〗　natrikso*K　〖Natrix　tigrina〗.　◆⏎

◎　代替可能の単語を示す、英字を囲む　〔　〕　は　半角の　［　］　に変えた

pulo.　※　～な背広　vulgara　［nerafinita］　jako.　※　こんな⏎

◎　見出しで、省略部分は（ ）で、代替部分は　((　))　にしています

◎　イタリックにはタグを付ける。　『　/　』の前後にはスペースを入れる

sin　　(re) iri　/　al　la　domo　/　hejmen.

◎　○に入った文字　㊟　は、〖注意〗　としている

◎　印刷間違いがあれば、後で検索して分かるように、〖●印刷間違い〗などと、〖●〗を付けて、補足してください

◎　その他、明確になっていないこともありますので、後半の説明も参照の上で、必要に応じ質問してください。

１．　これまでの経過：省略

２．　プログラム及び手作業で処理した主な内容は

スキャン画像から、ＯＣＲ認識によりテキストデータを作成し、最初に次の作業を行った。

但し、認識データが不備のため、**多くの見落としがあります**。

◎　次のような頁の区切り行を設定（L　R　は頁の左半分、右半分の記号）

pppp0031L##

◎　見出部分を　★　★　で囲み。頁左右内の通し番号と行番号を付加した

★ 10G32 いいん【医院】★　kuracejo;　kuracdomo.　◆⏎

（上から10番目の単語：　各頁左右で上から数えた単語の番号、G32　上から32行目にある）

◎　１頁の行数が左63行、右63行になるように、仮名の変わり目の空白行に
◎◎◎◎◎を入れた

◎　認識結果で全角ローマ字になっているのを半角ローマ字に変えるなど、文字の調整

◎　２個のスペースを１個にするなど、スペース（空白）の調整

◎　『　．　，　；　：　）　』等の後にスペースを入れ、
『　．　，　；　：　）　』等の前のスペースを削除する等。
『　／　』の前後にスペースを入れる

◎　印刷辞書での改行場所を示す記号　⏎　加える

◎　一つの見出データ（１行の場合も複数行の場合もある）の終了行の
最後に、記号　◆⏎　を加える。

◎　表示する場合に、見やすいように改行を行う為に、見出単語を使った
語句や、文例の前に、記号　※　を加える。

◎　専門語記号の〔〕を〖〗に変える

◎　PIV等を表す、単語末の、大文字『　I　』、『　K　』等を、インデックスで
あることを分かるようにする為、『　*I　』、『　*K　』等に変える。

◎　説明文の区分の数字等に、※〖1〗などと〖〗を付けて分かりやすくし、
後で改行出来るようにする

２．プログラムで上の変更を行う前の見出部分の読みあわせ修正を行う
ときに、前もって一緒に手作業で修正している部分。

①　動植物の学術名の　（　）　を　〖　〗に変える
これは、〖〗内の　x　文字を、字上符を表す　x　と区分する為。
字上符が表示出来無い場合用に、エスペラント単語で、字上符とx付与を
自動的に変換出来るようにする為

②　その他、特に気が付いた所。

③　この時点では、目で見て間違いを沢山発見出来るが、手を付けると
切りが無いので、気が付いても直していない。

日本語エスペラント辞典電子テキストデータ化仕様（案）

１．全体の基本構造

◎　ページが分かるように、下記のページの区切りを入れる
　pppp　に続けてページ番号を４桁の数字（前桁を０で埋める）　0031　等
　続けて　L　を左半分に、　R をは右半分に付ける

pppp0031L##

◎　１頁の行数が左63行、右63行になるように、仮名の変わり目の空白行に
◎◎◎◎◎　を入れた

★ 02G03 おんわ【温和】★　～な　milda;　pacama;　kvieta.　〔中庸⏎
の〕　modera.　◆⏎
◎◎◎◎◎⏎
◎◎◎◎◎⏎
◎◎◎◎◎⏎
◎◎◎◎◎⏎
★ 03G09 か【香】★　⇒　かおり．　◆⏎

従って、テキストデータは１頁で、必ず１２８行(左 1+63　右 1+63)となっている。

◎　見出部分を　★　ではさみ、１単語の基本構造は　★見出★　本文説明、例文等
◆⏎

★ 10G32 いいん【医院】★　kuracejo;　kuracdomo.　◆⏎

◎　一つの行の最後は必ず　改行マーク　⏎　で終わる。

◎　一つの単語（見出データ）は、　★　で始まり、　◆⏎　で終わる。

　ただし、説明が 1000 文字を超えているのは、途中にも見出しを入れ、番号を付す
(例)
★ 20G56 おも・う(1)【思う】★　※　〖１〗　〔思考〕　pensi;　略　⏎
〔考慮〕　konsideri.　〔意図〕　intenci.　〔信念〕　kredi.　⏎
　略・・・・・・・・・・
　略・・・・・・・・・・
　　略　※　君を　A　だと～った　Mi　prenis　vin　por　A.　◇⏎
★ 20G56 おも・う(2)【思う】★　※　〖４〗　〔期待〕　atendi.　略　⏎
※　～ったとおり　Kiel　mi　supozis　[atendis].　※　～ったよ⏎
　　略・・・・・・・・・・
　　略・・・・・・・・・・
gaŭ　pluvos.　◆⏎

２．見出部分の構造

○　見出し語の前にページの左右での掲載順の番号２桁、
『G』の文字の次に上からの行番号２桁を付加する。
基本形は　★ 99G99 ひらがな又はカタカナ【漢字】★
見出には次の形がある。
★ 08G11 いけぞんざい★
★ 08G27 いか・す★
★ 12G41 い・う【言う】★
★ 10G32 いいん【医院】★
★ 08G31 イエス★
★ 15G45 いカタル【胃－】★
★ 09G32 イエ(ズ)ス・キリスト★
★ 09G26 イカル(ガ)★
★ 13G31 いく-【幾】★
★ 11G35 きちじ((し))ょうてん【吉祥天】★
★ヒマラヤ（さんみゃく）【－（山脈)】★

●注意点　　接頭辞などの『　-　』、　　いく-
【　】内のカタカナの省略『－』　いカタル【胃－】　　胃カタル
カタカナ語の区分を示す『・』　　ア・ラ・カルト
動詞活用の区分点を示す『・』　　あい・する【愛する】
次の３行は、場合により行った（検討を要す）
省略可能を表すかっこ　（　）　イカル(ガ)　⇒　イカル、イカルガ
代替可能を表すかっこ　((　))　、　［　］　⇒　((　　))　に変えた
きちじ((し))ょうてん　⇒　きちじょうてん、きちしょうてん

３．説明部分での表記方法　（わかりやすいように、スペースは全角にした。）
◎　説明の分類の区分数字を〖〗で囲み、その前に改行記号　※　を付ける。
★ 27G56 おそ・い【遅い】★　※　〖１〗　〔時間〕　malfrua;　tarda.　⇒　おそ⏎
く．※　もう～いよ　Estas　jam　malfrue.　※　彼の帰りが⏎
～い　〔時刻〕　Li　revenos　malfrue.　〔なかなか帰ら⏎
ぬ〕　Li　ankoraŭ　ne　revenis.　※　彼の来るのが～かった　⏎
Li　venis　tro　malfrue.　※　〖２〗　〔のろい〕　malrapida;　lanta.　⏎
※　進歩が～い　La　progreso　iras　malrapide.　※　解りが～⏎
い　esti　lantkapa.　※　彼は何にかけても～い　En　ĉio　ajn　⏎
li　estas　tro　malrapida.　◆⏎

○　説明が　¶　マークで始まっている場合や、例文で始まっているとき、その前に改行記号　※　を付ける。
★ 20G60 おっとり★　※　～した　grandanima;　malurĝema;　kvieta;　⏎
senincita;　neincitiĝema.　※　～かまえる　sinteni　gran-⏎
danima　［malurĝema;　kvieta］.　◆⏎
★ 21G63 おっとりがたな【押取り刀】★　※　¶　～で駈けつける　alkuri　⏎

◎　派生語、例文で

○　例文、派生語の前に改行記号　※　を付ける。
○　その他、表示時に改行されていた方が見やすい所（●検討を要す）
（赤色下線部は間違いがあるので要修正箇所）

★ 21G37 いいだ・す【言い出す】★　〔口を切る〕　komenci　paroli;　⏎
ekparoli.　〔提案〕　proponi.　※　～しかねる　heziti　ekpa-⏎
roll.　※　君が～したんだよ　Neniu　a　五　a　krom　〔01　〕　vi ⏎
iiudatis.　※　～しべ　iniciatinto:　inidatoro.　◆⏎

4．説明部分での添え字を次のように変えて設定する
◎　PIV等を表す、単語末の、大文字『　I　』、『　K　』等を、インデックスであることを分かるようにする為、『　*I　』、『　*K　』等に変える。
この添え字は単語の頭にも有る。語頭の場合〖*I*〗等とする
例：　Ido*I.　　insinui*K.　　*I*gastrosuko.

5．専門分野記号の扱い。（説明部分の他の〔〕と区分する為）
◎　専門分野記号の〔〕を〖〗に変える
例：　〔植〕　⇒　〖植〗

6．動植物の学術名の　（　）　を　〖　〗に変える
これは、〖〗内の　x　文字を、字上符を表す　x　と区分する為。
字上符が表示出来無い場合用に、エスペラント単語で、字上符とx付与を自動的に変換出来るようにする為
例　（Rorippa　indica）　⇒　〖Rorippa　indica〗

7．説明部分での漢字、文字、記号とスペース

①　見出しの英字の場合で全角英字は基本的に半角に変えた。（●検討要す）
例：　イーピーばん【ＥＰ盤】　⇒　イーピーばん【EP 盤】

②　イタリック体の扱い　~~特に印を付けない~~
（紙の辞書を作成する時のためには必要だが、どうするか？）
最終的にタグを付けることにした 201605
◎　どちらか選択できる時のイタリック部分、『　/　』の前後にスペースを入れる
例：　(re) iri/*al*　*la*　*domo*/*hejmen*.
⇒　(re) iri　/　al　la　domo　/　hejmen.
◎　再帰代名詞のイタリック
例：　ne　senti　*sin*　viv (ant) a.　⇒　ne　senti　sin　viv (ant) a.

③　代替可能の単語を示す、英字を囲む　〔　〕　は　半角の　[　]　に変えた
説明文中の、補足説明の日本語と区分する為
amata　〔kara〕　infan (et) o.　⇒　amata　[kara]　infan (et) o.
（　★いどころ【居所】★　〔住所〕　loĝejo.　〔法的〕　domicilo.　）

④　その他 OCR 認識での間違い点に特に注意する必要あり

◎　漢字の数字　一　は　ー　やカタカナの長音　ー　と間違っている
　一　（コード 19968）は数字の一である、せんーコード 65293 ではない
　　カタカナの長音　ー　コード 12540 ではない
◎　・　は・にかえる
◎　ム　（コード 21430）はカタカナの「ム」でない
◎　数字のゼロ　0　と、英字の　O　o　との間違い
◎　o　は英字の小文字 o
◎　　カタカナの隣の　才　エ　タ　は　片仮名の　オ　　エ　タ　の間違い
◎　その他、色々有る

以下補足として、読みあわせで見つかるエラーの例

①　字上符の漏れ
　　例　prestiga　⇒　prestiĝa　　falsajo　⇒　falsaĵo

②　英字や字上符付き英字が漢字等別の文字に認識されたもの
　stari　が　⇒　starigi　　［ka　皿］　⇒　［kara］
　“彭　S　revido".　⇒　"Ĝis　revido".　　vo　と　⇒　voĉo
　励　⇒　eĉ　　e6　⇒　eĉ　　opo　吐　⇒　oponi.
　～免官〕aupeta　⇒　～免官　laŭpeta

③　英字の認識ミス　英字　で　ki　⇒ M　m ⇒ r　n　ci　⇒ d　ni ⇒ rd　等
　例　provoM　⇒　provoki　　domoŝelo ⇒ dornoŝelo　　solerd ⇒ soleni
　japaiaa　⇒　japana　　gastrokardno-　⇒　gastrokancino-　　sigrdfo　⇒　signifo
　　nerdu ⇒ neniu

④　イタリック英字のミス　　lafi　ぶ　ia,　peto　⇒　laŭ　*sia*　peto

⑤　添え字の認識漏れ（単語末の英字が小文字や別の ĵ になっている場合）
　例　ksenartroji　⇒　ksenartroj*I

⑥　ピリオドなどの間違い
　　al　iu,　⇒　al　iu.　　domagi　・　⇒　domaĝi.
　　animita　l　vigla　⇒　animita;　vigla

⑦　漢字の認識ミス
　　この仕事は m 大な～を持つ　⇒　この仕事は重大な～を持つ

⑧　文例中の区切りの所でピリオドが間違っている為、改行記号※が付かなかったもの
　　por　spiro　。　～もつかずに　⇒　por　spiro.　※　～もつかずに