

LINUX И ЯЗЫКИ ВОСТОЧНОЙ АЗИИ

С чего начать начинающему?

Под "языками Восточной Азии" в этой статье понимаются китайский, корейский и японский. В англоязычной документации применительно к ним распространён термин СЖК (сокращение от "Chinese, Japanese, Korean"). Вопросы компьютерной поддержки перечисленных языков сводятся в отдельную группу из-за своеобразия письменности, что требует и особого подхода в технической реализации. В последнее время к ним всё чаще относят вьетнамский язык (СЖКV), современное письмо которого основано на латинице, и язык ицзу (одной из китайских народностей), в котором употребляется пиктографическая письменность. "Добавления" объясняются тем, что проблема набора вьетнамского и ицзуского текста технически решена так же, как японского, китайского и корейского – отдельной программой, а не позицией в переключателе раскладок.

Статья знакомит читателя в первую очередь с технологиями набора текстов на китайском, японском и корейском языках. Как показывает практика, именно нетривиальность ввода иероглифики с клавиатуры ставит начинающего пользователя в тупик, и он порой не находит ничего лучшего, как поставить себе "китайский (японский, корейский) Windows". И это притом, что даже "в мире Windows" для этого вполне достаточно локализованной под вашу страну версии. Отчасти это объясняется тем, что в компьютерной литературе, издаваемой в России и на Украине, особенности работы с языками Восточной Азии абсолютно не рассматриваются и такое дело даже не упоминается. Пытаясь хоть в небольшой мере заполнить этот пробел, я постараюсь показать, что и Linux может предложить специалистам-восточникам широкие возможности для успешной работы.

Итак, что же нам нужно? Во-первых, "серьёзный" дистрибутив Linux, прежде всего из числа многодисковых и обеспечивающих многоязычную поддержку. Именно в составе таких дистрибутивов есть нужные вам приложения, а если какое-то из них вдруг отсутствует в "официальной поставке", вы быстро найдёте его в Интернете и без проблем установите. Во-вторых, правильно выбрать пакеты при установке дистрибутива. Выбирать следует всё, даже отдалённо связанное с вопросами локализации на нужные вам языки: тогда установятся и необходимые для работы "восточных язычных программ" системные библиотеки, шрифты, переводы интерфейсов и документация. В качестве раскладки клавиатуры по умолчанию (обязательно) и языка инсталляции (желательно) выбрать английский, а локализацию интерфейса настроить потом по вашему вкусу. Обратите внимание, что в отличие от английского, русского или немецкого языков, ввод китайского, японского и корейского нельзя настроить переключателем раскладок KDE, например.

Описание работы с языками Восточной Азии будет производиться на примере SuSE 9.1 Professional – именно этот дистрибутив имеет все необходимые приложения "из коробки", что и послужило для меня главным критерием выбора в его пользу. Все программы, о которых я буду рассказывать, шли прямо на компактках дистрибутива, и из Интернета на первых порах мне ничего не было нужно выкачивать. Кроме описанных в статье, есть и альтернативные "востокизаторы", о которых вы сможете узнать из документации или прямым поиском в Интернете.

Теперь несколько слов о том, откуда я всё это узнал. Первой мне попалась небольшая и легко написанная статья Джима Брина (Jim Breen) под названием "A Japanese Word-Processor for Linux", где автор даёт пример "японизации" своего OpenWriter'a из дистрибутива RedHat 7.x.

Оттуда же была ссылка на очень качественное руководство Майка Фабиана (Mike Fabian) по работе с языками Восточной Азии в Linux на примере дистрибутивов SuSE (<http://www.suse.de/~mfabian/suse-cjk.pdf>). Последняя версия (от 9 марта 2005 года) в формате pdf насчитывает 138 страниц. В этом документе хорошо объясняются теоретические основы СЖК в Linux, однако на китайский и корейский язык практических иллюстраций почти не приводится. На сайте ASP Linux одно время висела статья Павла Инкогнито "Японский в ASPLinux 9", но теперь я её там не нашёл. По корейскому языку есть "Ami Tutorial" Никлауса Гигера (Niklaus Giger), где рассматривается настройка кореизатора **ami** на примере Debian GNU/Linux. По китайскому же языку есть Chinese-HOWTO, украинский перевод которого (сделанный мною) находится на http://linuxsam.free.lafox.net/sxw/Chinese-HOWTO_ukr.zip (есть ещё doc- и pdf-версии). Разумеется, в Интернете вы можете найти и другие руководства, но мне было достаточно названных. Эта статья отнюдь не является переложением их содержания!

***Предостережение:** изложенная в статье информация не является руководством к действию и примером для подражания, а носит сугубо ориентировочный характер. За любой ущерб, нанесённый Вам, Вашему компьютеру и Вашему программному и аппаратному окружению попытками применить изложенные в статье рекомендации, автор настоящей статьи ответственности не несёт.*

Японский язык

Как известно, в японском языке используется смешанное письмо, состоящее из иероглифов и знаков слоговой азбуки. Эта последняя существует в двух видах – катакана и хирагана, которые оличаются сферой употребления и начертанием знаков. Кроме того, в японском языке находит применение и латинский шрифт – ромадзи. Японской латиницы существует несколько видов, из которых наиболее распространена хэпбёрновская транскрипция, придуманная для американцев и англичан, а потому ориентированная на орфографию английского языка. Хэпбёрновская ромадзи передаёт особенности японского произношения, а разные виды "собственно японской" латиницы (кунрэй-ромадзи) больше ориентированы на передачу латинскими буквами японского написания.

Если у вас стандартная клавиатура западного образца, то японский ввод работает следующим образом: набор знаков ромадзи – преобразование их в знаки слоговой азбуки – оформление иероглифического написания фразы.

kinput2 и canna

Для вызова японизатора наберите в эмуляторе терминала (Terminal Program) следующие команды:

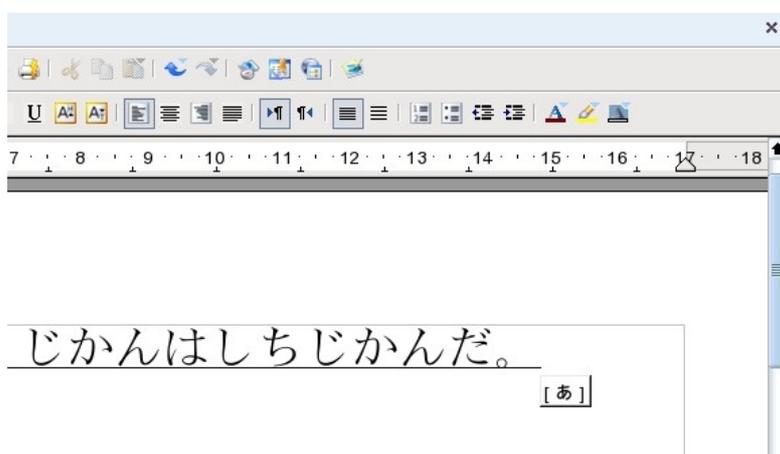
```
export LC_ALL=ja_JP.eucJP
export XMODIFIERS=@im=kinput2
kinput2 -xim -canna&OOo-writer
```

Вместо **eucJP** можно написать **UTF-8**, а вместо **OOo-writer** - подставить название любой другой программы, например **keddit** или **galeon**.

Японский ввод включается и отключается нажатием клавиши "пробел" при удержании клавиши Shift. Для некоторых приложений потребуется нажимать Ctrl+o. Если всё сделано правильно, возле курсора появится знак ввода хираганы – заключённая в квадратные скобки буква "А" этой слоговой азбуки. Теперь можно вводить японский текст, просто набирая его латинскими буквами:

koujoudehatarakujikannhashichijikannda.

Программа **kinput2** будет преобразовывать вводимую вами ромадзи в знаки хираганы. Обратите внимание, что показатель падежа тематического подлежащего (-wa) вводится через "ha", суффикс винительного падежа (-o) набирается как "wo", конечное "n" требует двойного нажатия соответствующей клавиши, а представление долгого [o] отражает японское написание азбукой, а не транскрипционную запись. Кроме того, если вы не в ладах с латиницей Хэпбёрна и больше любите кунрэй-ромадзи, можете сразу ею и пользоваться – японизатору одинаково.

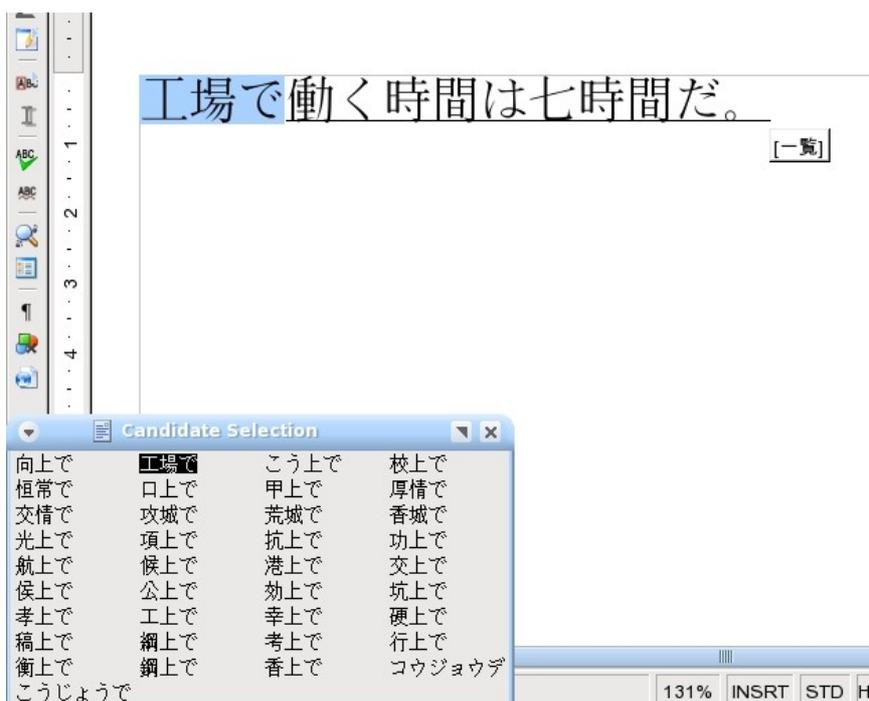


Если после того, как фраза набрана, вы нажмёте Enter, подчёркивание исчезнет и текст останется в таком виде. Конечно, употребление чистой хираганы в японском языке имеет место, но мы-то пишем "по-взрослому". Поэтому нажимаем "пробел", и приложение **canna**, анализируя последовательность набранных символов, предложит вам вариант иероглифической записи введённого текста.



Как видим, японизатор "ошибся" только в иероглифическом представлении первого слова. Он-то не может знать, какое из одинаково звучащих слов мы имели в виду, и подставляет то, которое до этого встречалось наиболее часто.

Нажимаем "пробел". Выскакивает окно выбора вариантов написания слова, в котором программа засомневалась. Для перехода между вариантами используются курсорные клавиши ("стрелки"), а закрепление выбора производится клавишей Enter.



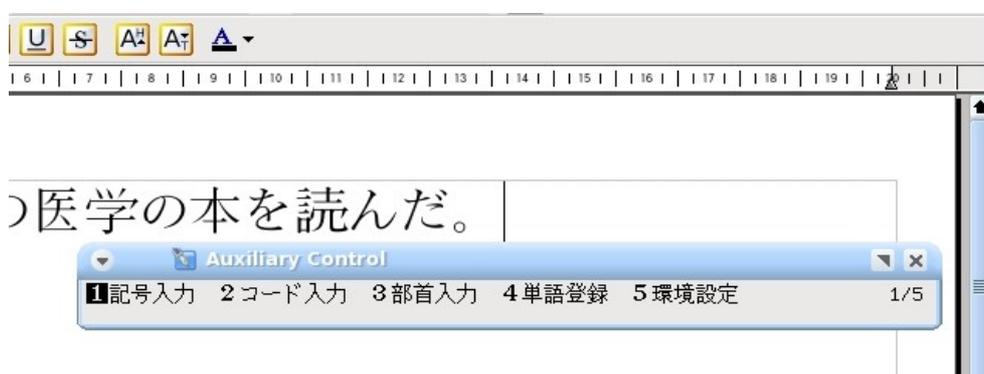
Если в качестве требуемой программы указать **kterm**, то можно будет создавать японские тексты прямо в эмуляторе терминала.



Теперь сделаем следующее. Наберите хираганой какое-нибудь слово, скажем nihongo, и, не трогая ни Enter, ни "пробел", понажимайте клавишу "стрелка вниз". Вы увидите, как слово меняет своё написание в такой последовательности: хирагана – полноширинная катакана – полуширинная катакана – полноширинная ромадзи – полуширинная ромадзи. Нажатия клавиши "стрелка вверх" прокручивают эти варианты в обратной последовательности. Таким способом удобно набирать гайрайго. Если вы, набрав хираганой какое-нибудь канго, нажмёте и отпустите "пробел", а потом будете нажимать "стрелку вниз", то увидите, как последовательно сменяются варианты написания данного слова. Выбрав нужный, вы можете закрепить его нажатием Enter, или, нажав "пробел" дважды, выйти в окно выбора вариантов написания вводимого слова.

Передвигать выделение по написанной фразе можно курсорными клавишами "вправо" и "влево". Например, слово hataraku из нашего примера вы решили изобразить катаканой. Тогда нажмите один раз "стрелку вправо", а потом – "стрелку вниз". Следующие её нажатия будут выводить разные варианты написания слова, а двойное нажатие "пробела" – список возможных вариантов его письменного представления (за исключением ромадзи и полуширинной катаканы).

Дополнительные возможности японизатора доступны по клавише Home (после нажатия Enter или до начала ввода японского текста). Появляется окно с пятью опциями: 1) ввод остальных знаков; 2) ввод цифровым кодом; 3) ввод иероглифов по радикалам; 4) управление словарями; 5) настройки программы. Ту или иную опцию можно выбрать нужной цифровой клавишей либо передвижением выделения в требуемую позицию (при помощи "стрелок" или последовательных нажатий "пробела") и нажатием Enter.



Пункты первый, четвёртый и пятый имеют свои подменю. Выбрав "ввод остальных знаков", вы сможете набирать различные символы, используемые в японской полиграфии, в том числе русские и греческие буквы и псевдографику. Опция номер четыре позволяет записать новое слово в словарь, удалить слово из словаря, а также подключать/отключать имеющиеся словари преобразования японского ввода в смешанное письмо. Пятая опция позволяет узнать текущие настройки программы и сообщает, где их можно изменить. "Ввод иероглифов по радикалам" полезен, когда вы не знаете, как читается тот или иной иероглиф. Выбрав необходимый радикал, вы получаете список иероглифов, в состав которых он входит, и дальше вам нужно только найти ("стрелками") и указать (нажатием Enter) требуемый символ.

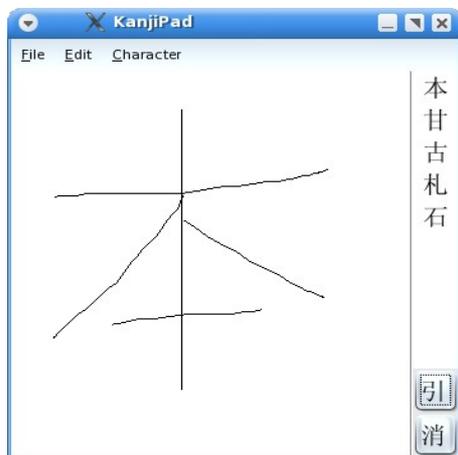
Выражение долготы гласного в катакане (для гайрайго) и в хирагане реализовано по-разному. Поэтому, набирая слова вроде sa:bisu или depa:to, вместо повторного нажатия клавиши "a" нужно нажимать "минус" верхнего цифрового ряда.

Закрывать дополнительные окна японизатора можно нажатием "пробела", если предварительно нажать и удерживать Shift.

OpenWriter позволяет снабжать иероглифику фуриганой. Для этого зайдите в меню Tools (Инструменты) и выберите опцию Ruby (Рубин).

kanjipad

Эта программка позволяет вводить иероглифы, "рисую" их мышкой. "Рисовать" иероглифы надо с соблюдением порядка черт, тогда желаемый символ будет в числе первых из предлагаемых, список которых выводится после нажатия кнопки "Принять". Щёлкнув по нужному иероглифу левой кнопкой мыши, заносим его в буфер, а потом вставляем куда следует щелчком средней клавиши мыши.



kiten

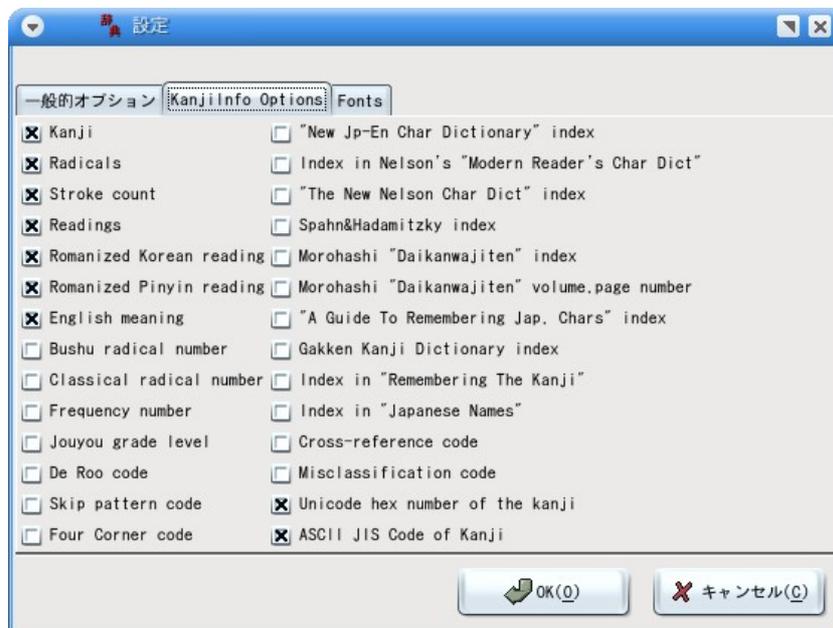
Это графический интерфейс двух японско-английских словарей – **edict** (словарь слов) и **kanjidic** (словарь иероглифов). Работает поиск и по английским словам из текста толкований и переводов. Иероглифы можно искать по радикалам и по количеству черт, при желании отфильтровывая редкоупотребительные. Программу **kiten** (как и **kanjipad**) не обязательно запускать под японской локалью: переключение ввода с латиницы на хирагану происходит по комбинации клавиш Shift+"пробел" или выбором из меню по правому щелчку мышью в строке поиска. Чтобы ввести знаки катаканы, во время набора удерживайте клавишу Shift.



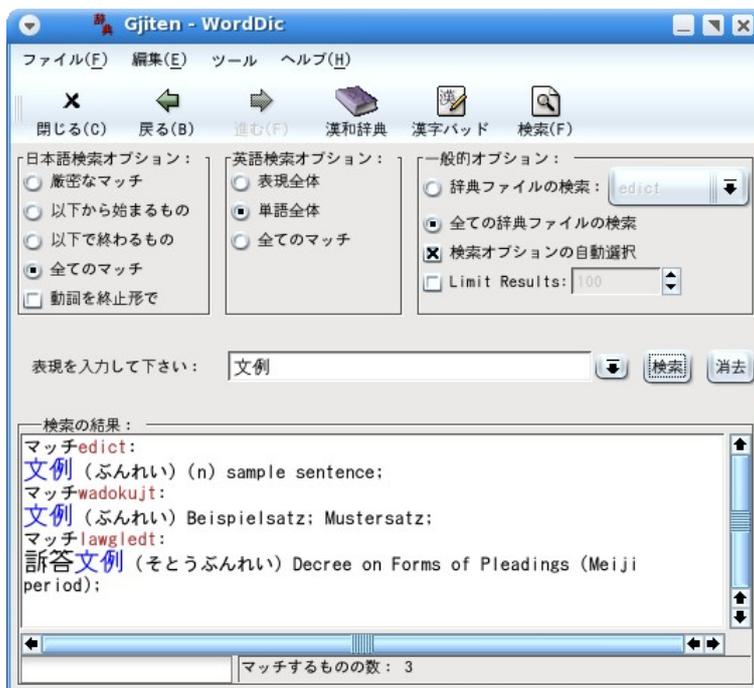
gjiten

Этой словарной оболочке не оказалось на дисках моего дистрибутива Linux (пришлось докачивать), поэтому сначала я не думал о ней писать. Однако программа обладает возможностями столь уникальными, что умалчивать о ней было бы свинством. В мире OpenSource эта программа занимает такое же место в инструментарии япониста, как The GIMP – в инструментарии вэб-дизайнера.

В отличие от **kiten'a**, **gjiten** нужно запускать под японской юникодной локалью UTF-8. В "штатную поставку" SuSE 9.1 Professional входят два уже упоминавшихся японско-английских словаря **edict** и **kanjidic**. Кроме них, ко **gjiten**'у можно подключить и другие словари того же формата, которые можно скачать из Интернета. Среди них большинство специализированных, но особого уважения заслуживает полный японско-немецкий словарь "Вадоку-дзитэн" – более 20 мегабайтов чистого текста. Для иллюстрации полноты сведений, выводимых этой программой, просто приведу скриншот:



Пример поиска слова:



И пример поиска иероглифа:



Эта программа интегрируется с **kanjipad**'ом, в котором нужно писать требуемый иероглиф, если вы хотите в строку поиска ввести его прямо, или же есть возможность найти его в словаре по ключу и количеству черт (учтено даже то обстоятельство, что число черт у некоторых иероглифов спорно). В общем, качайте – не пожалеете.

Корейский язык

В корейском языке тоже принято смешанное письмо – слоговая азбука хангыль и китайские иероглифы ханджа.

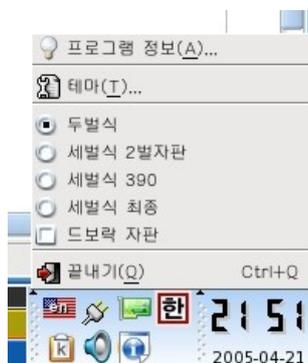
nabi

Для запуска **nabi** наберите в эмуляторе терминала (Terminal Program) такие команды:

```
export LC_ALL=ko_KR.eucKR
XMODIFIERS="@im=nabi"
export XMODIFIERS
nabi&OOo-writer
```

Как и в случае с японским (и с китайским) языками, здесь тоже можно изменить кодировку и название нужной программы. Если всё сделано правильно и вы работаете в KDE, программа заявит о себе индикатором её состояния в системном лотке. Возможны три состояния индикатора: ввод латиницы, ввод хангылей и путешествие по опциям основной программы (т.е. в данном случае OpenWriter'a) с помощью клавиатуры. Переключаться между состояниями "хангыль/латиница" можно комбинацией клавиш Shift+"пробел", а чтобы войти

в "третье состояние", нажмите F10 (работает не во всех программах). В **nabi** есть возможность настроить внешний вид индикатора и выбрать корейскую раскладку клавиш из нескольких вариантов (правый щелчок мышью по индикатору).



Для преобразования только что набранного (он подсвечивается) хангыля в ханджа нажмите F9. Появится список иероглифов, отвечающих данному чтению. Некоторые комбинации хангылей отданы под ввод других типографских знаков, список которых появляется аналогично. Выбирать нужный знак следует нажатием цифровой клавиши или перемещаясь по их списку "стрелками" "вправо-влево", закрепляя выбор нажатием Enter. Список высвечивает десять первых "кандидатов", а прокручивать "серии кандидатов" можно нажатием "пробела" или "стрелок" "вверх-вниз".



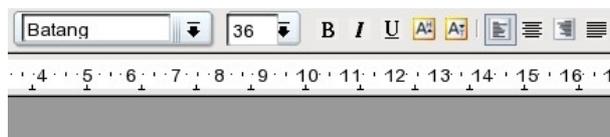
ami

Для запуска **ami** дайте в эмуляторе терминала (Terminal Program) такие команды:

```
export LC_ALL=ko_KR.eucKR
XMODIFIERS=@im=Ami
ami -xim&OOo-writer
```

Запущенный корейзатор возвещает своё появление звуковым сигналом и сворачивается на панель задач. Включать ввод корейского языка можно комбинацией Shift+"пробел". Впрочем, **ami** обладает развитым графическим интерфейсом, позволяющим делать настройки без необходимости править конфигурационные файлы. Преобразование хангыля в ханджа здесь аналогично **nabi**, отличается только вид всплывающего окна да добавлены опции желаемого представления знаков ханджа. Обратите внимание, что клавиша Enter служит для закрепления введённого слова и перехода на новую строку, а для простого

закрепления набранных символов служит комбинация Ctrl+Enter. Но так это по умолчанию, и вы легко сможете пере назначить комбинации клавиш, "подняв" окно корейзатора с панели задач и пройдясь по вкладкам графического интерфейса.



По клавише F3 доступен набор символов, напрямую не связанных с корейским языком, в том числе ввод русских и греческих букв, многобуквенных сокращений, трактуемых каждое как отдельный символ, и знаков японской слоговой азбуки.



Перемещаться по списку символов можно курсорными клавишами, а закреплять выбор – нажатием Enter'a или "пробела". Перебор вкладок – клавишей Tab за исключением нескольких последних (русские и японские символы), которые выбираются щелчком мыши.

Кстати, зайдя в меню Tools (Инструменты) OpenWriter'a, вы можете обнаружить пункт Hangul/Hanja Conversion (Преобразование хангыль-ханджа). В версии OpenOffice'a для Linux этот пункт не работает до OpenOffice.org 2.0beta включительно, а в версии для Windows заработал начиная с упомянутого выпуска.

Китайский язык

В современном китайском письме используется два вида иероглифов – полного (традиционного) и упрощённого написания, в связи с чем в компьютерном наборе китайского текста различают "два китайских языка", точнее, два вида письма – китайское

упрощённое (Simplified Chinese) и китайское традиционное (Traditional Chinese). Хочу подчеркнуть, что здесь речь не идёт об "упрощённом" и "традиционном" китайском языке, и не о двух видах клавиатурных раскладок, а лишь о характере письма.

Исторически сложилось, что пионерами "китаизации" Linux выступили жители о.Тайвань, где, во-первых, приняты иероглифы традиционного написания, во-вторых, есть свой взгляд на обозначения тонов в китайских слогах, и в-третьих, чувствуется японское влияние.

В отличие от японского и корейского языков, где способ ввода, можно считать, один (если отбросить варианты японских национальных клавиатур и разных корейских раскладок), для китайского языка существует множество способов ввода, причём едва ли не каждый из них имеет разновидности. Большинство китаистов России и стран бывшего СССР, да и многие северные китайцы владеют только одним способом ввода – пиньинем, так как он почти не требует обучения именно набору – достаточно знать китайский язык.

Все способы ввода, за исключением "рисования" мышкой, распознавания по голосу и цифровой кодировки, можно грубо разделить на две группы – фонетические и графические. Первые сродни рассмотренному выше способу ввода японского текста, а вторые олицетворяют попытку составить иероглиф из его элементов, закреплённых за определёнными клавишами на клавиатуре компьютера (это чем-то отдалённо похоже на ввод корейских хангылей).

Проблемам китайского ввода в Интернете посвящено не много, а очень много материала, пересказать который здесь при всём желании невозможно. Поэтому для тех, кто знакомится с этим вопросом впервые, я изложу только принцип, взяв за основу китаизатор **xcin** и способ набора пиньинь.

Традиционное письмо

Наберите в эмуляторе терминала:

```
export LC_ALL=zh_TW.big5
XMODIFIERS="@im=xcin"
export XMODIFIERS
xcin&OOo-writer
```

После того, как китаизатор запустится, проверьте его работоспособность, нажимая Shift+"пробел" (выбор между полноширинным и полуширинным стилем письма) или левые Ctrl+Shift (выбор способа ввода). Поскольку эти сведения выносятся в окно китаизатора, изменение информации вы сразу заметите.

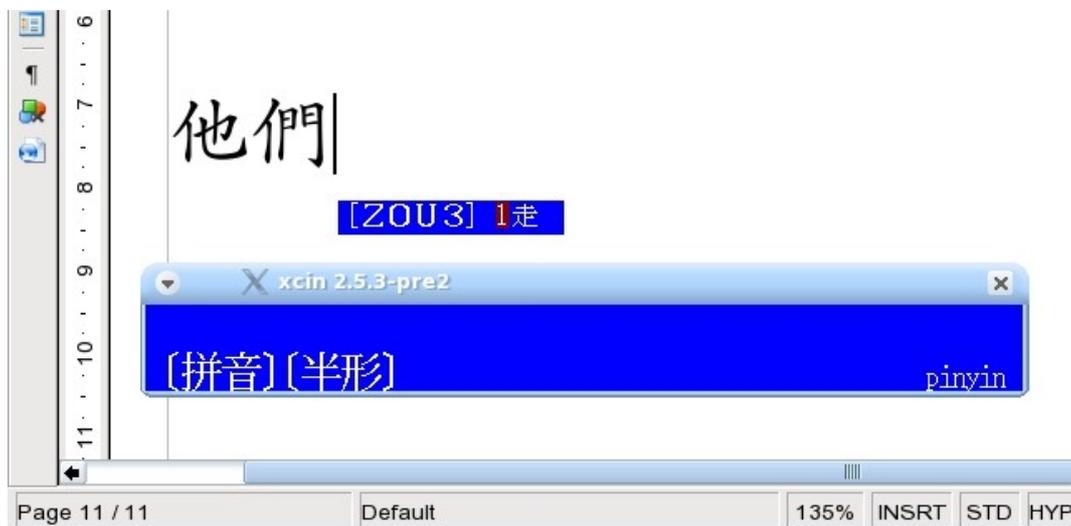


В китайской типографике принято, что под каждый символ отводится знакоместо одинаковой ширины (при этом ширина самого знака может быть разной). Знаки препинания можно переносить на следующую строку. Поэтому если желательно, чтобы ваш текст выглядел "истинно по-китайски", используйте полноширинный режим.

Изначально китаизатор настроен на ввод английского текста и арабских цифр. Переключаясь между способами ввода, можно выбрать zh_hex (шестнадцатеричный код), sj (цанцзе), simplex (модернизированный цанцзе), phone (ввод азбукой чжунь-цзыму), jyutping (фонетический ввод в произношении диалекта юэ), bimsphone (модернизированный phone), bimsponyin (модернизированный пиньинь), agra30 (рекомендуется пользователям **GNU emacs**), jyutping0 (усовершенствованный jyutping), pinyin (почти "классический" пиньинь).

"Ключевыми клавишами" китайского ввода являются "пробел", цифры верхнего клавиатурного ряда и "стрелка вниз".

При наборе пиньинем тоны китайских слогов обозначаются цифрами после буквенного выражения слога. Однако на Тайване принят иной порядок обозначения тонов. Для первого тона никакой цифры вводить не надо. А нейтральный тон как раз и обозначается цифрой 1. Итак, для примера напишем китайский эквивалент фразы "Они заходят в парк" (по пекинской номенклатуре ta1menzou3jin4gong1yuan2). Выбираем способ ввода pinyin и нажимаем клавиши (подчерк обозначает "пробел"): "ta_1men1_zou3_jin4_lgong_2yuan2_4" (естественно, без кавычек). Цифры после пробелов обозначают номер иероглифа в предлагаемом списке. Чтобы набрать "точку", нужно нажать "точку" цифровой части клавиатуры. Эта точка имеет "европейский вид", что допустимо на Тайване. Чтобы ставить "китайскую" точку, вам придётся воспользоваться вставкой символа другим способом, например, из меню OpenWriter'a или автозаменой после набора всего текста. К сожалению, при данном способе ввода вам так же точно придётся вводить каплевидную запятую и другие знаки препинания. Кроме того, я не нашёл, как вводить слоги вроде lü и nü – приходится пользоваться опцией вставки специального символа из меню текстового процессора или переключаться на другой способ ввода.



Этих недостатков почти лишён способ набора bimsponyin. Вот последовательность нажатия клавиш для ввода той же фразы (вертикальная черта обозначает "стрелку вниз", подчерк – "пробел", всё без кавычек): "ta_|2_men1_zou3_jin4_gong_|2_yuan2|4_". Чтобы ввести "китайскую точку", нужно нажать Shift+"точка американской раскладки". Аналогично вводятся и другие китайские знаки препинания (в том числе кавычки), кроме каплевидной запятой. Слоги lü и nü вводятся как lv и nv соответственно. При других способах ввода (и в других китаизаторах) слоги с [ü] могут быть в конце списка слогов с [u].

Если в правой части полосы списка иероглифов появляются "уголки", это значит, что количество иероглифов, соответствующее данному чтению, не умещается в отображаемой части списка и вам предлагается прокрутить его курсорными клавишами или "пробелом".

Упрощённое письмо

Наберите в эмуляторе терминала:

```
export LC_ALL=zh_CN.GB2312
XMODIFIERS="@im=xcin-zh_CN.GB2312"
export XMODIFIERS
xcin&OOo-writer
```

Для иероглифов упрощённого написания **xcin** предлагает другие способы ввода, но пиньинь остаётся и здесь. Присутствуют zh_hex, pinyin, shuangpin (быстрый фонетический способ – количество нажатий на клавиши уменьшается примерно вдвое против пиньиня), jtcj_gb (разновидность цанцзе), phonegb2 (набор азбукой чжуинь-цзыму), wubizixing (насколько я могу судить, самый популярный и "профессиональный" из графических способов – он позволяет очень точно "выцепить" нужный знак без необходимости пролистывать длинный список).

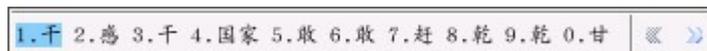
В отличие от традиционного написания, пиньинь упрощённого значительно более прост. Обозначения тонов вводить не нужно, а звуки [ü] и [u] не различаются. Поэтому списки иероглифов - "кандидатов" для ввода в набираемый текст - здесь длиннее. И опять-таки проблема ввода знаков препинания, например вопросительного. Для их набора можно переключаться на английский язык (средствами самого китаизатора).

scim

Это мощнейший "востокизатор", позволяющий набирать китайско-корейско-японско-латинские тексты "без пересадок". Есть развитый графический интерфейс, в котором можно тонко настроить поведение программы. Работа с "востокизатором" во многом аналогична работе с приложениями, описанными выше. Приведу только скриншоты:



Это панель управления способом ввода – в данном случае цзыжаньма. Управление производится мышью или комбинациями клавиш.



Как вы догадываетесь, это – список иероглифов-кандидатов. Для перемещения по нему служат клавиши "вверх" и "вниз", а ввод осуществляется "пробелом" или номерной цифрой.



В этом окне видна часть предоставляемых способов ввода. Их действующий комплект зависит от локали, в которой запущена программа.

OpenOffice.org 2.0beta (точнее, 1.9.79) имеет встроенный преобразователь иероглифов упрощённого написания в иероглифы полного и наоборот. Выход на него производится через меню Tools (Инструменты) – Language (Язык) - Chinese translation (Преобразование китайского). Однако имейте в виду, что такие преобразователи делают свою работу с ошибками, вплоть до искажения смысла всего предложения. Так что слепо на них полагаться не следует. Удачи!

Статья свободна для некоммерческого распространения и использования. Коммерческое распространение и использование данной статьи – только по договорённости с автором.

Винник Валерий aka Lao,
wwalerijn_na_mail_toqka_ru

2005-04-30