NATIONAL RESEARCH UNIVERSITY

Лингвистические электронные ресурсы для лингвистических исследований (введение)

О.Н.Ляшевская
школа лингвистики НИУ ВШЭ

На какие ресурсы нужно опираться при исследовании (неизвестного) языка?

Что нужно для исследования языка?

Документация языков:

можно ли написать текст на незнакомом языке, прочитав его **грамматику** и **словарь**?

Что нужно для исследования языка?

Документация языков:

можно ли написать текст на незнакомом языке, прочитав его **грамматику** и **словарь**?

Нет - нужно прочитать много **текстов**, а еще лучше, пообщаться с носителями языка.

- *язык - средство общения*
- *в языке всегда есть много вариантов выражения мысли - в зависимости от намерений говорящего и коммуникативной ситуации*
- *язык - живой, языковые средства могут меняться*

Цифровая революция и лингвистика

книги
радио, телевидение
словари
энциклопедии
учебники

+

электронные книги
аудиокниги
интернет
электронные словари
электронные пособия,
медиакурсы, тренажеры
автоматические переводчики,
Skype-обучение ....

Язык в Интернете

(Ру)нет как фонд текстов на (русском) языке:

• Источник полезной информации: новости, статьи из газет и журналов, электронные версии книг, сценарии кинофильмов, сайты музеев и учебных заведений, обзоры товаров, транскрипты интервью...

+ аудио- и видеозаписи: радиопрограммы, интервью, аудиокниги, радиоспектакли, песни, youtube (rutube)

• Сети для электронной коммуникации: facebook, вконтакте, Живой журнал и др., форумы, чаты

• Справочные, энциклопедические и образовательные ресурсы

• Поисковые системы и переводчики

Интернет представляет испорченный язык

(Ру)нет как фонд текстов на (русском) языке:

• Источник полезной информации: новости, статьи из газет и журналов, электронные версии книг, сценарии кинофильмов, сайты музеев и учебных заведений, обзоры товаров, транскрипты интервью...

+ аудио- и видеозаписи: радиопрограммы, интервью, аудиокниги, радиоспектакли, песни, youtube (rutube)

• Сети для электронной коммуникации: facebook, вконтакте, Живой журнал и др., форумы, чаты

• Справочные, энциклопедические и образовательные ресурсы

• Поисковые системы и переводчики

Электронные библиотеки (примеры)

Google Books - books.google.com

- **"Народные" проекты**
- lib.ru Библиотека Максима Мошкова
- lib.aldebaran.ru Библиотека "Альдебаран"
- netslova.ru Сетевая словесность
- russ.ru Русский журнал

- **Академические проекты**
- feb-web.ru Фундаментальная электронная библиотека "Русская литература и фольклор" -- аннотированные электронные версии классики, включая варианты изданий (там же словари и литературные энциклопедии)

- **ru.wikipedia.org Википедия (архив Википедии как большой текстовый ресурс)**

- Проект Гутенберг - www.gutenberg.org

Электронные корпуса

• Задачам лингвистического исследования лучше всего отвечают не просто **тексты** (архивы текстов),

а **корпуса**

- коллекции текстов, снабженные специальной разметкой (информация о текстах в общем, о каждом предложении и слове)

Электронные корпуса

• Типичные вопросы, на которые отвечают корпуса:

- отличается ли речь авторов-женщин от авторов-мужчин?

<все тексты должны иметь помету "пол автора">

- когда впервые появилось в языке слово *слямзить*? (NB! не появилось, а задокументировано)

<все тексты должны иметь помету "дата создания">

<корпус должен уметь находить слово во всех формах - разметка лексем>

- отличается ли сочетаемость слов *хотеть* и *стремиться*? (ср. ?*я стремился, чтобы...*)

Классификация ресурсов

- грамматики
- корпуса
- словари

- базы данных

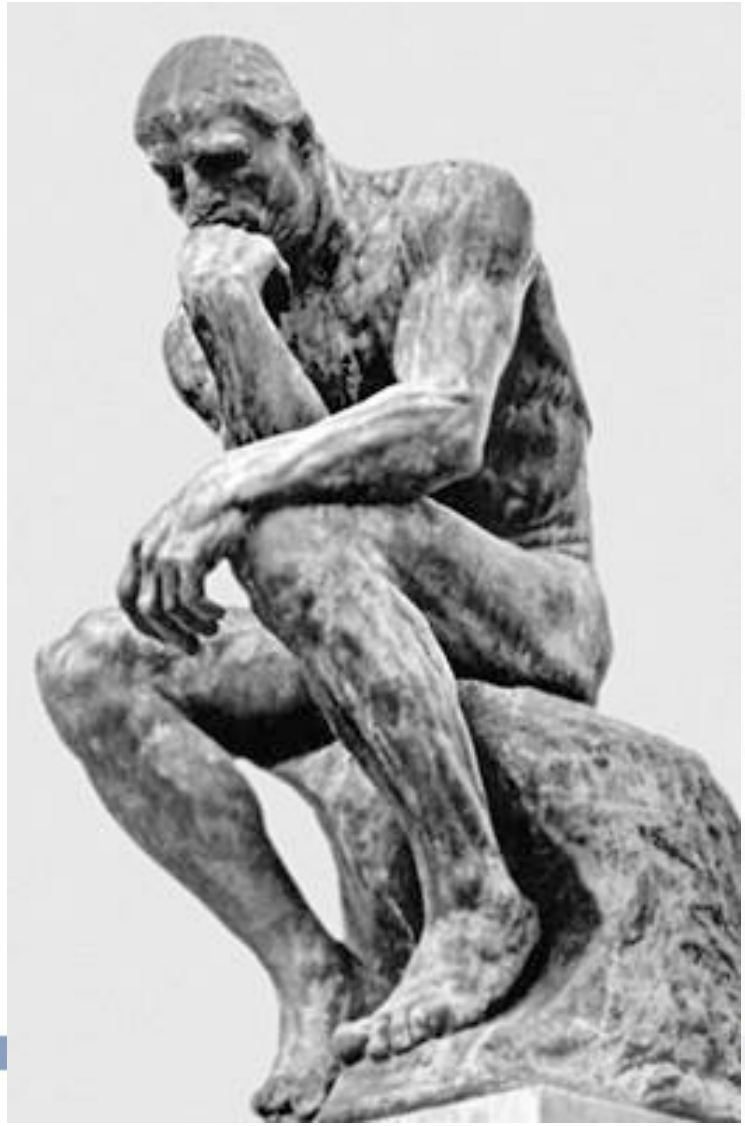
- структурированные факты по грамматикам

- сводные данные по употреблению языковых единиц в корпусе, в т.ч. частотные

- структурированные факты о лексике

- справочные системы
- другие специальные ресурсы (геоинформация по диалектам и т.п.)

Чем еще пользуются лингвисты?



Чем еще пользуются лингвисты?

- интуицией:

если являешься носителем языка, можно спросить себя, "можно ли так сказать"?

Чем еще пользуются лингвисты?

- интуицией:

если являешься носителем языка, можно спросить себя, "можно ли так сказать"?

К сожалению, интуицию трудно превратить в ресурс!

Чем еще пользуются лингвисты?

- интуицией:

если являешься носителем языка, можно спросить себя, "можно ли так сказать"?

- опросами информантов + экспериментами

если сомневаешься, можно спросить носителя языка, "можно ли так сказать"?

можно (в ходе эксперимента) спровоцировать носителя языка произнести или не произнести интересующую меня языковую единицу

К сожалению, результаты опросов информантов и экспериментов малодоступны и еще не стали общественным достоянием

Примеры (just a few...)

Словари в электронном формате

- **dic.academic.ru**
  - словари и энциклопедии на русском языке
- **slovari.ru** под эгидой Института русского языка им. В.В. Виноградова
  - академические словари русского языка + грамматики русского языка
- **feb-web.ru/feb/feb/dict.htm** - словари Фундаментальной е-библиотеки
- **etymolog.ruslang.ru** - этимологические словари
- **dict.ruslang.ru** - словари на основе НКРЯ
- **ru.wiktionary.org** - Вики-словарь (сделай словарь сам!)
- **http://www.onelook.com** - поиск по словарям английского языка
- ~~**slovari.yandex.ru**~~ на портале Яндекса

и др. словари русского языка – двуязычные словари – энциклопедии онлайн

Электронные корпуса русского языка

- **ruscorpora.ru** - Национальный корпус русского языка
- **ruTenTen** (sketch engine)
- **корпуса Сергея Шарова** (Лидс, Англия)
- **Упсальский корпус**
- **Тюбингенский корпус**
- **ХАНКО** Хельсинкский аннотированный корпус русских текстов
- **Компьютерный корпус текстов русских газет конца XX века** (МГУ)
- **Корпус русского литературного языка** (С.-Петербург)
- **Регенсбургский диахронический корпус русского языка** (древнерусские тексты)
- Рукописные памятники Древней Руси: берестяные грамоты, летописи, рукописная книга
- Параллельный корпус переводов «Слова о полку Игореве»
- Корпус русских публицистических текстов второй половины XIX века

Электронные корпуса других языков

- BNC - Британский национальный корпус (английский язык)
- corpus.byu.edu/coca - корпус COCA (American English)
- COHA - исторический корпус американского английского
- COSMAS - Мангеймские корпуса немецкого языка
- Lexicum - корпус канадского французского
- VISL - корпуса Университета Южной Дании (Э.Бик)
- Leeds corpora - интернет-корпуса Университета Лидса (С.Шаров)
- web-corpora.net - корпуса осетинского, калмыцкого, бурятского, монгольского, албанского и многих других "небольших" языков
- JRC-Acquis - параллельный корпус документов Евросоюза на 23 языках
- см. больше ссылок на studiorum.ruscorpora.ru и linguistilist.org

Базы данных

О языках (типологические БД)

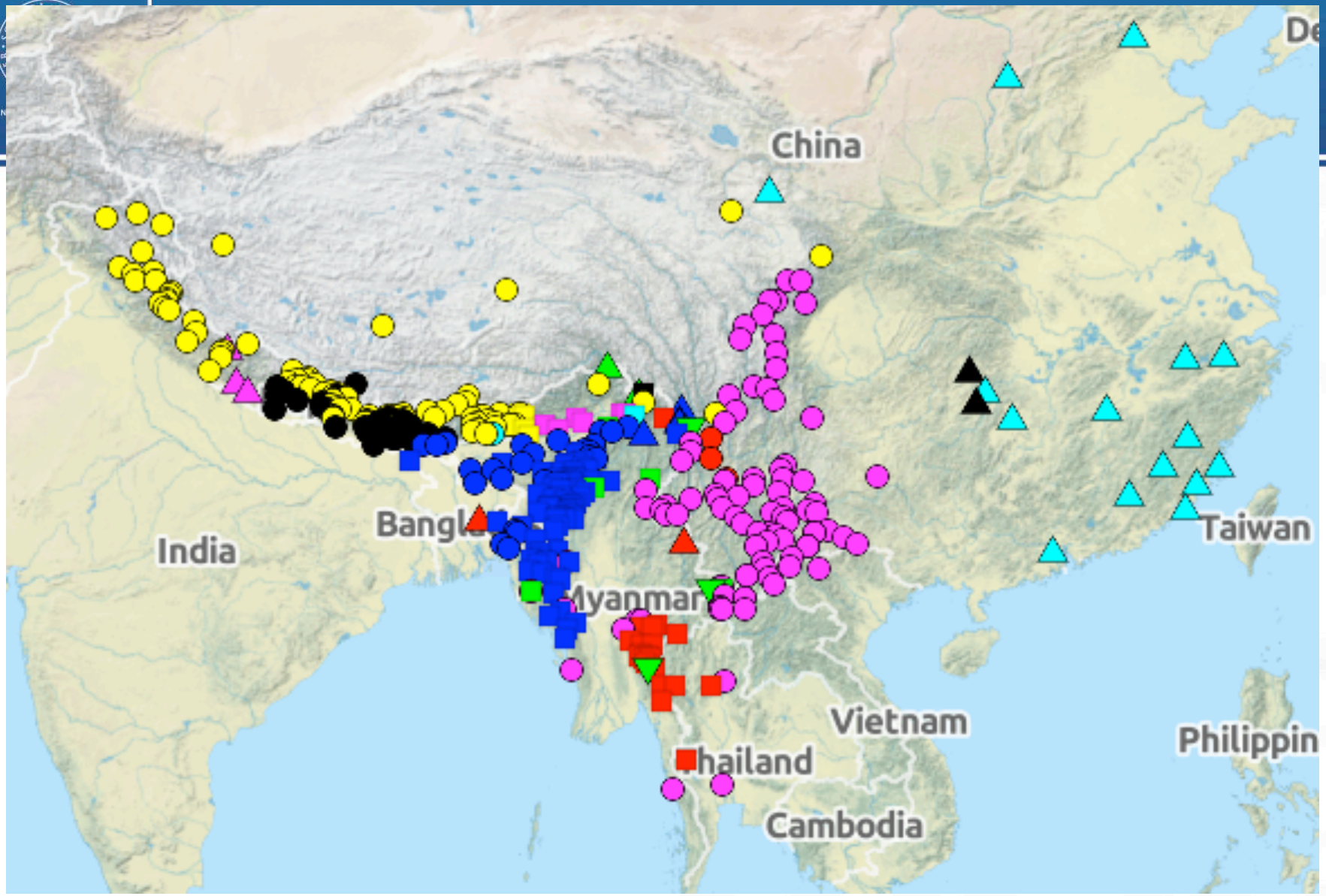
- **Ethnologue** www.ethnologue.com - база данных языков мира (семья, численность, ареал, живой/вымирающий, карты)
- **Glottolog** glottolog.org - генеалогическая классификация + библиография
- **WALS** wals.info - The World Atlas of Language Structures, + типологические свойства языков, типологические очерки

О лексике

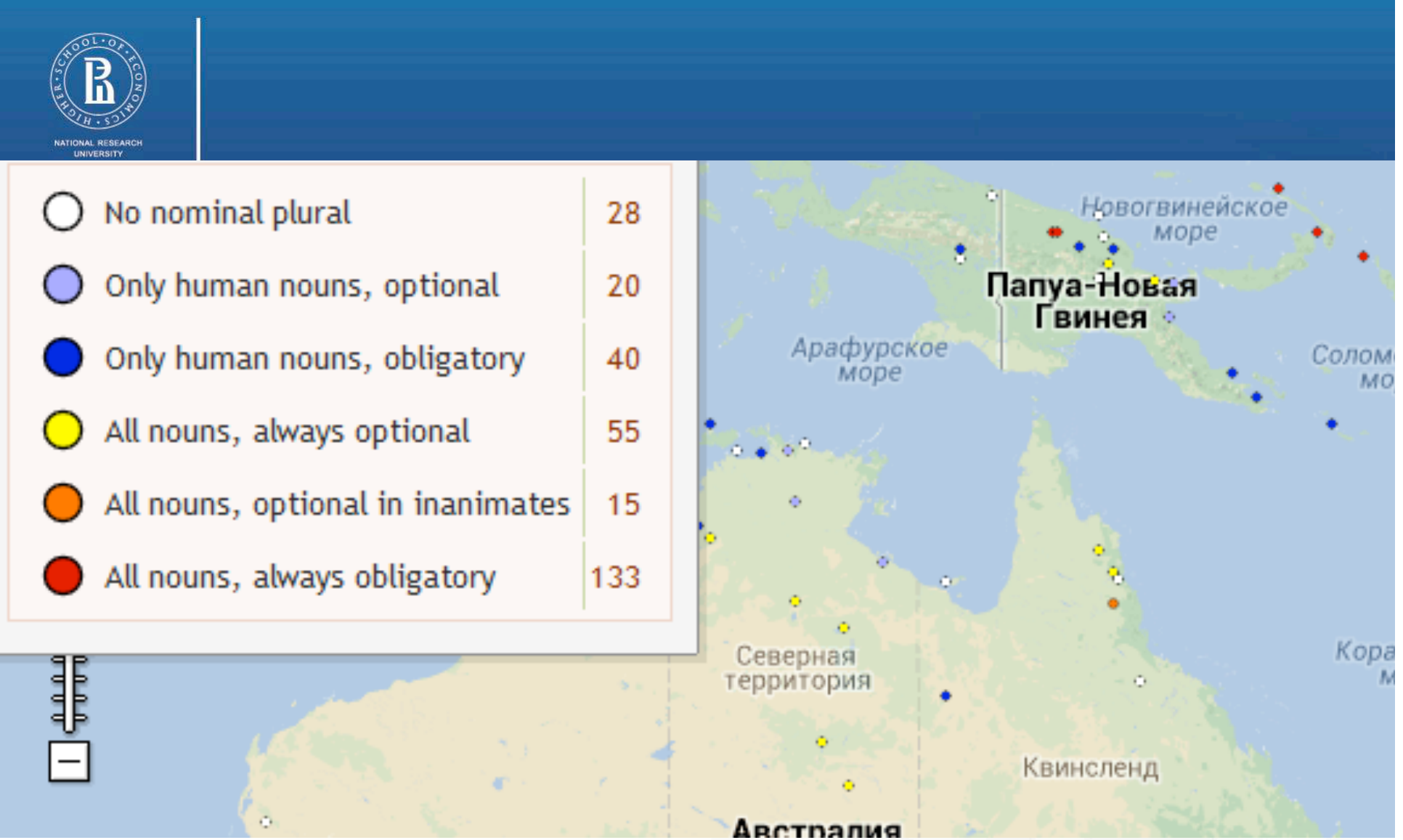
- **The Tower of Babel** starling.rinet.ru - этимологическая база данных
- **WordNet** - семантическая сеть лексики (синонимы, гипонимы...)
- **visuwords.com** - online graphical dictionary, связи между словами

О синтаксисе и сочетаемости

- **SketchEngine** - помощник лексикографа
- **FrameNet** - конструкции глаголов



Glottolog: ареал распространения сино-тибетских языков



WALS: языки с разными системами грамматического числа

Number: 1713
Proto: *wasa
English meaning: calf, deer calf
German meaning: Kalb, Renkalb
Finnish: vasa 'Kalb, einjähriges Renkalb', vasikka 'Kalb'
Estonian: vasik, vasikas (gen. vasika)
Saam (Lapp): vyesi (I), vi̯̊sse (T), vūi̯ss (Kld.), vuai̯ss (Not.) 'klein
Mordovian: vaz (E M), vazńe (E), vazńä (M) 'Kalb'
Mansi (Vogul): (wę̄səj KM, wę̄səɣ P, wāsiɣ So. 'Elchkalb' - rejec

Вавилонская башня: этимологически связанное гнездо (фино-угорск.)

goal *(noun)* **ukWaC freq = 168345** (107.5 per million)

object_of	58924	3.2
score	8390	11.28
achieve	9422	9.9
concede	1421	9.39
accomplish	585	7.97
reach	1924	7.66
net	337	7.42
pursue	648	7.41
attain	400	7.35
grab	406	7.34
set	2413	7.01
pull	501	6.88
disallow	190	6.67

subject_of	25451	2.4
score	903	8.59
disallow	223	8.04
concede	204	7.53
gape	76	6.5
come	1316	5.44
kick	76	5.44
rule	61	5.24
orientate	34	5.06
arrive	90	4.43
cap	20	4.38
beat	53	4.31
direct	53	4.22

modifier	67879	1.6
ultimate	1911	9.27
long-term	875	7.66
league	638	7.38
winning	401	7.33
primary	993	7.24
second	2000	7.19
common	1529	7.17
strategic	645	7.1
realistic	422	7.05
achievable	290	6.97
stated	259	6.8
score	611	6.75

modifies	11026	0.3
scorer	389	9.39
kick	634	8.86
tally	129	7.9
keeper	204	7.31
scramble	50	6.75
drought	78	6.65
difference	676	6.28
cushion	53	6.26
lead	267	6.24
setting	405	6.14
kicker	25	6.04
post	482	5.91

SketchEngine: типичные контексты слова *goal*

clever/intelligent

ukWaC freqs = 20589/26115

clever 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 intelligent

and/or	4955	10062	2.2	3.6
perceptive	0	34	0.0	6.4
thought-provoking	0	32	0.0	6.2
informed	0	66	0.0	6.2
autonomous	0	46	0.0	6.2
adaptive	0	39	0.0	6.1
well-informed	0	24	0.0	6.0
literate	0	26	0.0	5.9
compassionate	0	27	0.0	5.9
well-educated	0	17	0.0	5.7
cultured	0	19	0.0	5.7
rational	0	46	0.0	5.6
playful	0	22	0.0	5.6
sensitive	8	134	2.0	5.9
thoughtful	14	121	5.0	7.7
affectionate	6	31	4.5	6.2
sophisticated	23	75	4.2	5.7
charming	21	50	4.9	5.9
insightful	11	31	5.2	6.1
resourceful	12	29	5.8	6.3
witty	132	166	8.3	8.2
inventive	22	24	5.9	5.5
clever	54	30	5.8	4.8
funny	233	103	7.0	5.7
cunning	16	7	5.9	4.0
catchy	19	0	5.8	0.0

modifier	4950	3168	0.9	0.5
emotionally	0	111	0.0	8.6
artificially	0	52	0.0	7.9
fiercely	0	26	0.0	7.0
highly	0	570	0.0	6.9
ferociously	0	8	0.0	6.2
supposedly	0	28	0.0	6.2
averagely	0	7	0.0	6.1
moderately	0	11	0.0	5.7
reasonably	0	54	0.0	5.7
computationally	0	6	0.0	5.6
supremely	0	7	0.0	5.5
culturally	0	12	0.0	5.5
exceptionally	29	25	6.0	5.9
remarkably	24	11	5.7	4.8
amazingly	17	7	5.9	5.0
wonderfully	20	9	5.4	4.5
very	1707	596	5.6	4.0
too	476	76	5.4	2.8
damn	12	0	5.6	0.0
dead	16	0	5.8	0.0
diabolically	9	0	5.9	0.0
awfully	15	0	6.1	0.0
terribly	25	0	6.2	0.0
devilishly	17	0	6.8	0.0
fiendishly	45	0	8.1	0.0

modifies	10948	16081	2.0	2.4
being	0	208	0.0	6.1
robot	0	77	0.0	6.1
agent	9	455	0.4	6.0
guess	0	35	0.0	5.5
routing	0	27	0.0	5.3
layman	0	22	0.0	5.3
conversation	0	88	0.0	5.1
creature	11	137	2.4	5.9
lyric	81	80	5.8	5.7
fellow	52	14	5.1	3.1
pass	67	9	5.2	2.2
stuff	146	6	5.1	0.4
gimmick	15	0	5.1	0.0
satire	19	0	5.1	0.0
flick	21	0	5.2	0.0
lob	15	0	5.3	0.0
pun	19	0	5.3	0.0
ruse	17	0	5.5	0.0
eh	24	0	5.8	0.0
wordplay	21	0	5.8	0.0
chap	47	0	5.9	0.0
twist	94	0	6.5	0.0
trick	166	0	6.7	0.0
clog	50	0	7.0	0.0
ploy	68	0	7.2	0.0

SketchEngine: синонимы и сочетаемость слов *clever* и *intelligent*

Справочно-информационные ресурсы

- **Справочные порталы** (на примере русского)
- **gramota.ru** - Грамота.ру, портал "Русский язык"
  - Справочная служба русского языка
  - словари, статьи, интерактивные диктанты, игры
- **gramma.ru** "Культура письменной речи"
  - академические словари русского языка
  - грамматики русского языка
- **pishu-pravilno.livejournal.com** "Пишу правильно", сообщество в Живом журнале

- **Порталы для исследователей**
- **studiorum.ruscorpora.ru** - справочная система в помощь пользователям корпусов
- **linguistlist.org** - информация о конференциях, журналах, исследователях и институтах, ресурсах и т.д.

Ресурсы компьютерной лингвистики

- Яндекс.ru, Google.com, Bing.com, Baidu.com - работают на гигантских размеченных (индексированных) архивах текстов и на специальных словарях

- Системы проверки орфографии - используют словари и базы данных

- Системы автоматического перевода (translate.google.com, multitran и другие) - используют параллельные корпуса и словари

- Системы классификации новостей - словари + базы знаний

- Системы анализа мнения о товарах (opinion mining) и т.д.

Инструменты для разметки текстов и создания ресурсов

Аннотация текстов:

- ELAN tla.mpi.nl/tools/tla-tools/elan/ - аннотация аудио и видео
- Praat www.fon.hum.uva.nl/praat/ - для работы с фонетикой
- GATE GATE.ac.uk - профессиональные инструменты
- UIMA uima.apache.org компьютерной лингвистики

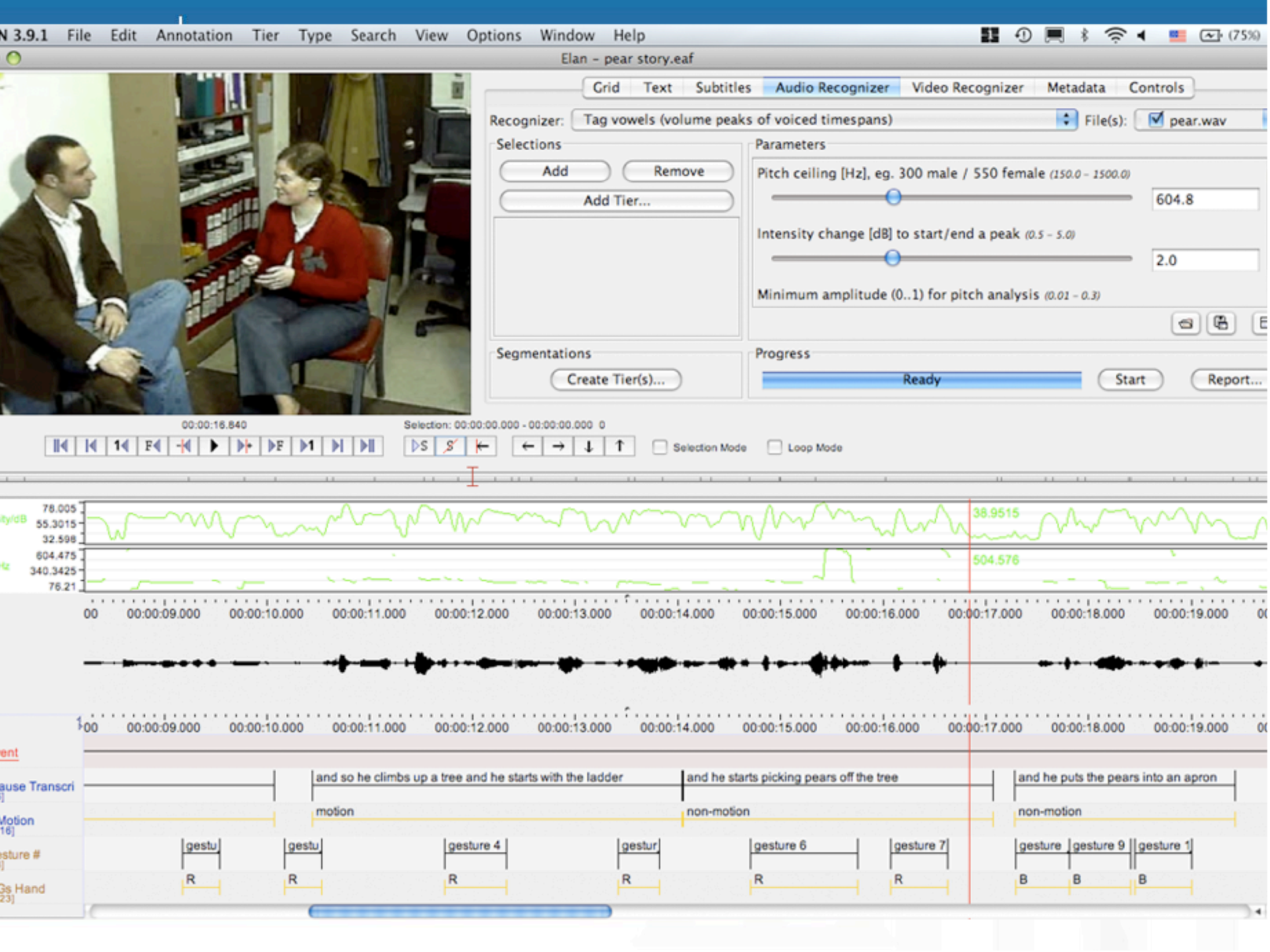
Создание словарей:

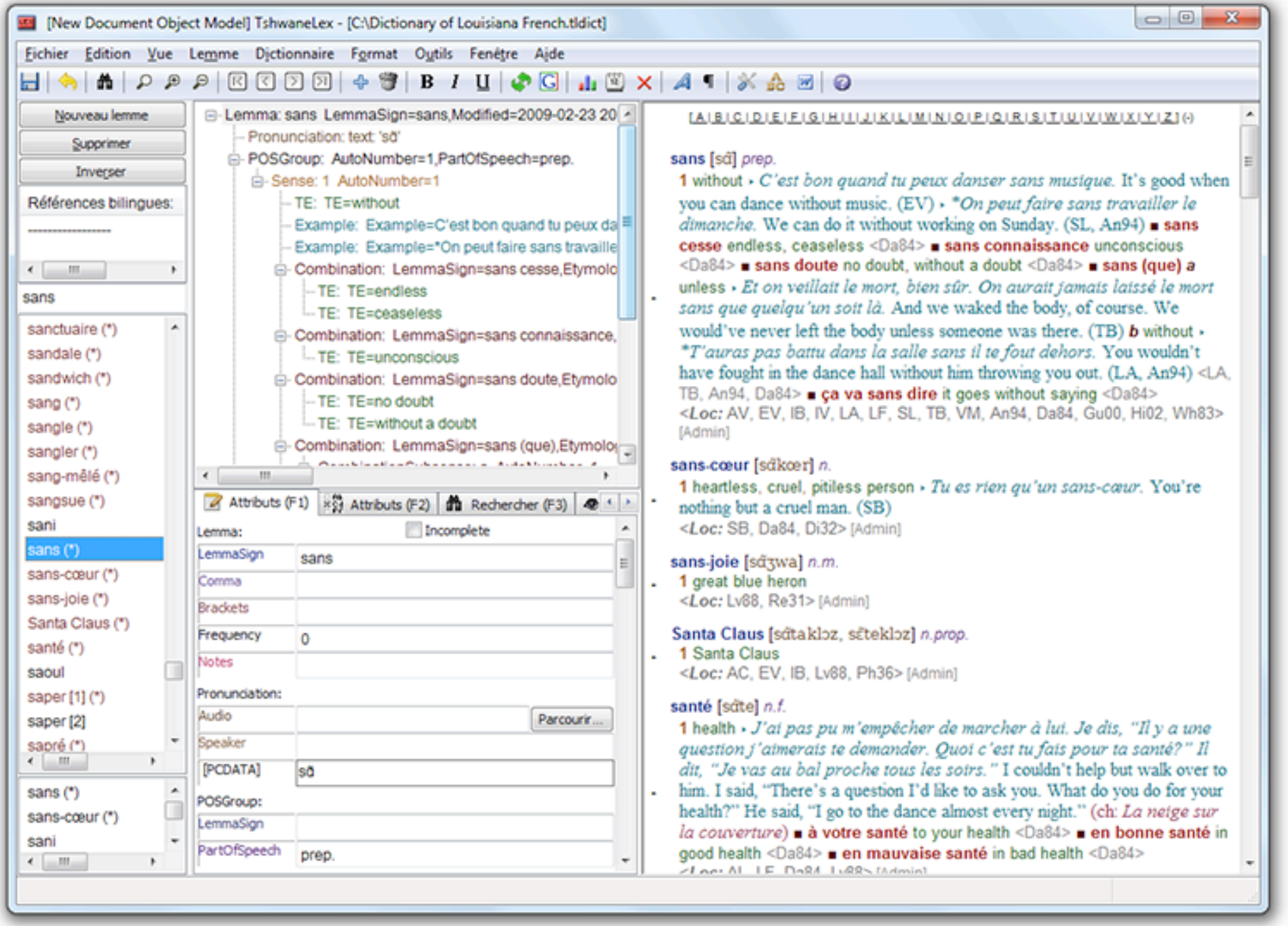
- Lexus tla.mpi.nl/.../lexus/
- iLex www.emp.dk
- IDM idm.fr
- TshwaneLex tshwanedje.com/tshwanelex/
- Lexique Pro lexiquepro.com
- ABBYY Lingvo Content

Корпус-менеджеры:

- WordSmithTools
- Bonito
- Corpus Workbench

- + текстовые редакторы, базы данных, конкордансеры и т.д


N 3.9.1 File Edit Annotation Tier Type Search View Options Window Help
Elan - pear story.eaf
Grid Text Subtitles Audio Recognizer Video Recognizer Metadata Controls
Recognizer: Tag vowels (volume peaks of voiced timespans)
File(s): pear.wav
Selections
Add Remove
Add Tier...
Parameters
Pitch ceiling [Hz], eg. 300 male / 550 female (150.0 - 1500.0)
604.8
Intensity change [dB] to start/end a peak (0.5 - 5.0)
2.0
Minimum amplitude (0..1) for pitch analysis (0.01 - 0.3)
Segmentations
Create Tier(s)...
Progress
Ready
Start
Report...
00:00:16.840
Selection: 00:00:00.000 - 00:00:00.000 0
Selection Mode
Loop Mode
78.005
55.3015
32.598
38.9515
604.475
340.3425
76.21
504.576
00:00:09.000
00:00:10.000
00:00:11.000
00:00:12.000
00:00:13.000
00:00:14.000
00:00:15.000
00:00:16.000
00:00:17.000
00:00:18.000
00:00:19.000
and so he climbs up a tree and he starts with the ladder
and he starts picking pears off the tree
and he puts the pears into an apron
motion
non-motion
non-motion
gestu
gestu
gesture 4
gestur
gesture 6
gesture 7
gesture
gesture 9
gesture 1
R
B


[New Document Object Model] TshwaneLex - [C:\Dictionary of Louisiana French.tldict]
Fichier Edition Vue Lemme Dictionnaire Format Outils Fenêtre Aide
Nouveau lemme
Supprimer
Inverser
Références bilingues:
sans
sanctuaire (*)
sandale (*)
sandwich (*)
sang (*)
sangle (*)
sangler (*)
sang-mêlé (*)
sangsue (*)
sani
sans (*)
sans-cœur (*)
sans-joie (*)
Santa Claus (*)
santé (*)
saoul
saper [1] (*)
saper [2]
sapré (*)
Lemma: sans LemmaSign=sans,Modified=2009-02-23 20
Pronunciation: text: 'sɑ̃'
POSGroup: AutoNumber=1,PartOfSpeech=prep.
Sense: 1 AutoNumber=1
TE: TE=without
Example: Example=C'est bon quand tu peux da
Example: Example=*On peut faire sans travaille
Combination: LemmaSign=sans cesse,Etymolo
TE: TE=endless
TE: TE=ceaseless
Combination: LemmaSign=sans connaissance,
TE: TE=unconscious
Combination: LemmaSign=sans doute,Etymolo
TE: TE=no doubt
TE: TE=without a doubt
Combination: LemmaSign=sans (que),Etymolo
Attributs (F1)
Attributs (F2)
Rechercher (F3)
Lemma:
Incomplete
LemmaSign
sans
Comma
Brackets
Frequency
0
Notes
Pronunciation:
Audio
Parcourir...
Speaker
[PCDATA]
sɑ̃
POSGroup:
LemmaSign
PartOfSpeech
prep.
sans (*)
sans-cœur (*)
sani
[A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z]
sans [sɑ̃] prep.
1 without ▸ C'est bon quand tu peux danser sans musique. It's good when you can dance without music. (EV) ▸ *On peut faire sans travailler le dimanche. We can do it without working on Sunday. (SL, An94) ■ sans cesse endless, ceaseless <Da84> ■ sans connaissance unconscious <Da84> ■ sans doute no doubt, without a doubt <Da84> ■ sans (que) a unless ▸ Et on veillait le mort, bien sûr. On aurait jamais laissé le mort sans que quelqu'un soit là. And we waked the body, of course. We would've never left the body unless someone was there. (TB) b without ▸ *T'auras pas battu dans la salle sans il te fout dehors. You wouldn't have fought in the dance hall without him throwing you out. (LA, An94) <LA, TB, An94, Da84> ■ ça va sans dire it goes without saying <Da84>
<Loc: AV, EV, IB, IV, LA, LF, SL, TB, VM, An94, Da84, Gu00, Hi02, Wh83>
[Admin]
sans-cœur [sɑ̃kœr] n.
1 heartless, cruel, pitiless person ▸ Tu es rien qu'un sans-cœur. You're nothing but a cruel man. (SB)
<Loc: SB, Da84, Di32> [Admin]
sans-joie [sɑ̃ʒwa] n.m.
1 great blue heron
<Loc: Lv88, Re31> [Admin]
Santa Claus [sɑ̃takloz, sɑ̃teklɔz] n.prop.
1 Santa Claus
<Loc: AC, EV, IB, Lv88, Ph36> [Admin]
santé [sɑ̃te] n.f.
1 health ▸ J'ai pas pu m'empêcher de marcher à lui. Je dis, "Il y a une question j'aimerais te demander. Quoi c'est tu fais pour ta santé?" Il dit, "Je vas au bal proche tous les soirs." I couldn't help but walk over to him. I said, "There's a question I'd like to ask you. What do you do for your health?" He said, "I go to the dance almost every night." (ch: La neige sur la couverture) ■ à votre santé to your health <Da84> ■ en bonne santé in good health <Da84> ■ en mauvaise santé in bad health <Da84>

Инструменты для разметки текстов и создания ресурсов

Аннотация текстов:

- ELAN tla.mpi.nl/tools/tla-tools/elan/ - аннотация аудио и видео
- Praat www.fon.hum.uva.nl/praat/ - для работы с фонетикой
- GATE GATE.ac.uk - профессиональные инструменты
- UIMA uima.apache.org компьютерной лингвистики

Создание словарей:

- Lexus tla.mpi.nl/.../lexus/
- iLex www.emp.dk
- IDM idm.fr
- TshwaneLex tshwanedje.com/tshwanelex/
- Lexique Pro lexiquepro.com
- ABBYY Lingvo Content

Корпус-менеджеры:

- WordSmithTools
- Bonito
- Corpus Workbench

- + текстовые редакторы, базы данных, конкордансеры и т.д