

Počítačová lexikografie Makrostruktura

Adam Rambousek

Lexikografické podklady

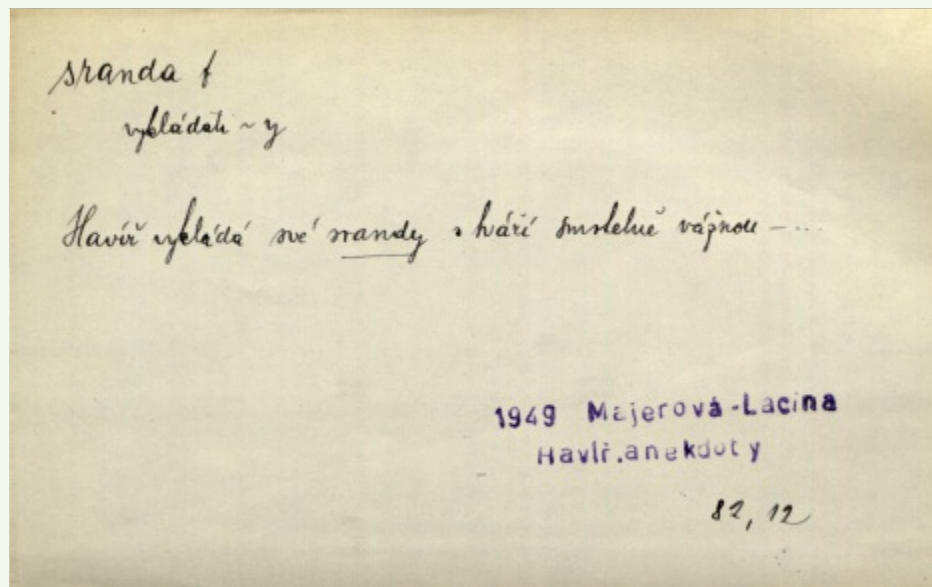
- důkazy o použití jazyka
 - intuice
 - excerpta, výpisky
 - korpusy
- intuice (armchair linguistics)

Intuice

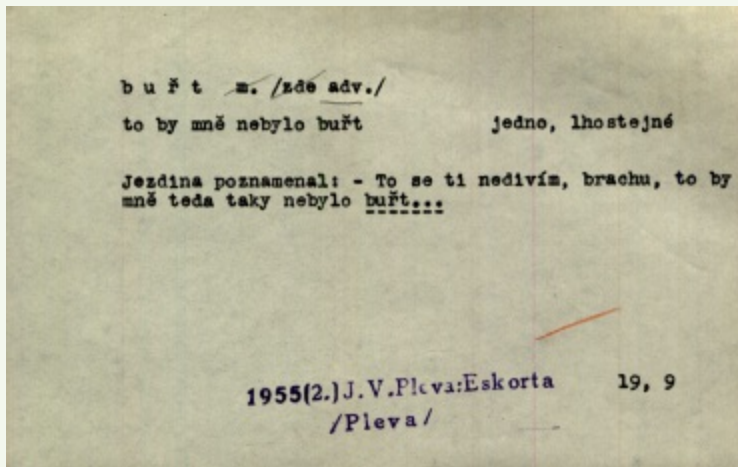
- In the absence of objective evidence, introspection was appealed to instead. But studies in corpus linguistics have shown that introspection is a very flawed technique. We human beings are wired to register the unusual in our minds, generally in a way that is available to conscious recall. But we fail to pay any attention to the commonplace patterns of usage on which we rely so heavily in our everyday communications. Patrick Hanks (Euralex 2000)
- Should it ever come about that linguistics can be carried out without the intervention and suffering of a native speaker analyst, I will probably lose interest in the enterprise.
Charles Fillmore ("Corpus linguistics" or "Computer aided armchair linguistics")

Výpisky

- Appeal to the English speaking and English reading public, 1879
- Návod pro sběratele materiálu k "Slovníku jazyka českého", 1911
 - 8 696 850 lístků (1911 - 1991), neologismy 270 538 záznamů



Výpisky



slo		Bleskový filtr	
slo	myš	Slovní druh	podst ž
esilí	myš (polohovací to zařízení)		
lek	Sbohem Simpsonovi!	Rubrika	Život a vůbec
Autor	Ivan Straka	Postavení autora	redaktor
Mluví		Postavení mluvčího	
Kontext	Na redakční poradě bylo odbojným redaktorům nařízeno rozkazem, že musím používat v souvislosti s myší sloveso "poklepat" (kliknout či cvaknout v žádném případě). Byla šance oživit českou terminologii, ať už fonetickým bastardem "kliknout" či ryze českým buřetáckým "cvaknout". Zvítězilo absurdní "poklepat", ač činnost, kterou s myši (polohovacím to zařízením) provádíme, má k poklepání tak daleko jako Mohamed k hoře.		
Poznámka			
Zdroj	Softwarové noviny	Datum	16.10.2013
Číslo	10	Strana	132
		Rok	1995
Tweet 0		Error	
		Exportovat	

Výpisky

- výhody
 - posuny významu
 - terminologie
 - šíření lexikografie
- nevýhody
 - pracné, časově náročné
 - subjektivní (časté výjimky)

Korpus

- IB047 Úvod do korpusové lingvistiky
- a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language as a source of data for linguistic research
- dokonalý korpus neexistuje
 - korpus je jen vzorek jazyka
 - obsahuje i nespisovný jazyk
 - čas a náklady na výrobu

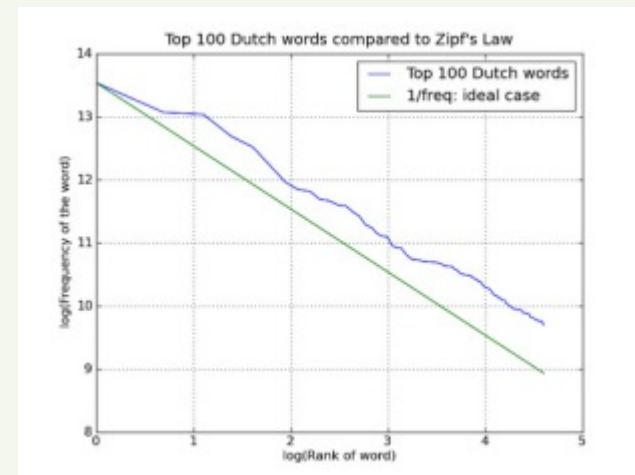
Korpus

➤ velikost

- Brown Corpus (1960) milion slov (10^6)
- COBUILD (1980) 20 milionů slov (10^7)
- BNC (1990) 100 milionů slov (10^8)
- OEC (2000) miliarda slov (10^9)
- TenTen 10^{10} slov

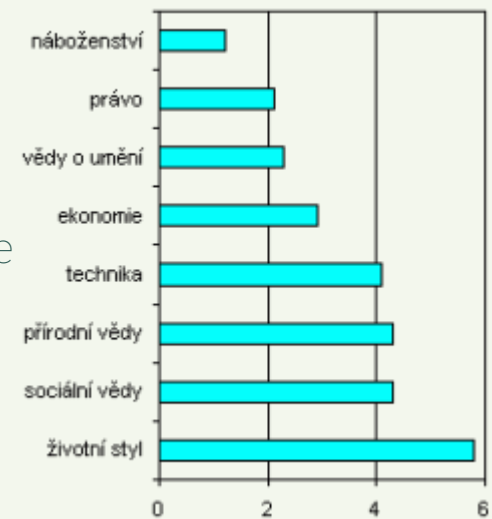
➤ Zipfův zákon (1935) několik slov s vysokou frekvencí, mnoho slov s nízkou frekvencí

- 10. slovo je 10x častější než 100. slovo



Korpus

- vyvážený
 - Linguistic Data Consortium – anglický korpus z článků Associated Press a New York Times
 - BNC – Journal of Gastroenterology, mucosa x unfortunate
- co zahrnout a v jakém poměru?
- BNC
 - 90% written, 10% spoken; 75% informative, 25% imaginative
- SYN2000 (100 milionů slov)
 - 60% publicistika, 25% odborná, 15% beletrie
- SYN2005 (100 milionů slov)
 - 40% beletrie, 27% odborná, 33% publicistika



Korpus

- získání
- převod
- značkování (formální, lingvistické)
- z webu: Sketch Engine, WebBootCaT

Lexikální databáze

- podrobná strukturovaná jazyková databáze
 - (nyní obvykle) doklady z korpusu
 - gramatické údaje
 - valence, vzory
 - styl, užití, oblast...
 - vztahy mezi slovy
- podklad pro slovníky a výzkum
- PraLeD (Pražská Lexikální Databáze)
- DANTE (Database of ANalysed Texts of English)

Lexikální databáze

game

in a structured, essentially non-physical, activity involving one or more people, esp one engaged in for enjoyment or to pass the time

COLLOCATE TYPE OBJECT OF **COLLOCATES** play

- Then, with wind howling round the windows and rattling the doors, we sat in front of a peat fire and played **games** with the children.
- *Bullseye* (07 February 2006) A great **game** on TV but a pretty awful DVD game.

STRUCTURE PP_X of

- It follows that the **game** of chess, in its effects upon mental character, is greatly misunderstood.
- England and Dublin at the time had started the **game** of bingo and the idea reached Graiguecullen in the mid '40s.
- The oldest game still played seriously in pubs is undoubtedly the **game** of darts.

CHUNK game of chance

- Description: Sic Bo, meaning dice pair, is an ancient Chinese **game of chance** played with three dice.
- His application for the chair of anatomy and botany was decided by drawing of lots and he was unlucky in this **game of chance**.

STRUCTURE N_mod

COLLOCATE TYPE EQUIPMENT **COLLOCATES** card, board, table, pen-and-paper

- Countess Spencer often stayed at Park House while Ruth, Lady Fermoy taught the children card **games**.
- At one time most games relevant to history were board **games**.
- For social interaction, painting, drawing, and **table games** such as dominoes and cards.
- The new "games area" includes two for memory development (" Pairs " and " Copy Cat "), a simple game called " BreakIt " which most children will be familiar.

COLLOCATE TYPE PLACE/OCCASION OF PLAYING **COLLOCATES** casino, party, parlour, pub

- Blackjack is by far our most popular casino **game**.
- Maybe some of our older readers can remember an old party **game** "hunt the thimble".
- The pursuer will also arrange activities during the day and evening, ranging from quizzes and parlour **games** to deck tennis or c.
- These include pub **games** against opposing teams to determine a winning side.

COLLOCATE TYPE CONTENT/DESCRIPTION **COLLOCATES** puzzle, guessing

- A puzzle **game** scrambles puzzle pieces for children to unscramble.
- Basically it is a guessing **game** involving coins.

COLLOCATE TYPE PARTICIPANTS **COLLOCATES** panel

- [TV-RAD] → David Baddiel has devised a new Radio 4 panel **game**, to be recorded next month.
- [TV-RAD] → Acknowledgement: This game is a simplified version of a UK TV **quiz game** called " The Weakest Link ".

LINK ahead of the game **LINK** all part of the game **LINK** anybody's game **LINK** beat sb at their own game **LINK** give the game away **LB** **LINK** the only game in town **LINK** game show **LINK** games soon **LINK** game theory

The screenshot shows a web-based lexical database interface for the word "jazyk". The main entry is for "jazyk" (language), with a note that it is a homonym. The interface includes a search bar, a list of grammatical categories (VI, E1, E2, E3, E4, E5), and a detailed description of the word's usage and morphology. The word is listed as "jazyk" (masculine, singular, nominative case). The description includes: "Přívod deževce: SSJČ svaňatý, velmi pohyblivý orgán v dutině ústní (u zvířat v tlamě, zobáku atd.); orgán chuti, sluchu". The grammatical categories are: E1 k VI Adj+SUBST (rozbalit/sbalit), E2 k VI SUBST+Adj. (rozbalit/sbalit), E3 k VI SUBST+Subst-gen (rozbal), E4 k VI Subst+SUBST-gen (rozbal), and E5 k VI SUBST+Prop+Subst/SUBST+Subst-ji. The interface also includes a search bar, a list of grammatical categories, and a detailed description of the word's usage and morphology. The word is listed as "jazyk" (masculine, singular, nominative case). The description includes: "Přívod deževce: SSJČ svaňatý, velmi pohyblivý orgán v dutině ústní (u zvířat v tlamě, zobáku atd.); orgán chuti, sluchu". The grammatical categories are: E1 k VI Adj+SUBST (rozbalit/sbalit), E2 k VI SUBST+Adj. (rozbalit/sbalit), E3 k VI SUBST+Subst-gen (rozbal), E4 k VI Subst+SUBST-gen (rozbal), and E5 k VI SUBST+Prop+Subst/SUBST+Subst-ji. The interface also includes a search bar, a list of grammatical categories, and a detailed description of the word's usage and morphology.

Makrostruktura

- heslář (+předmluva, přílohy...)
- heslo¹ = lemma, entry term, heslové slovo, headword
 - obvykle nominativ sg., slovesa v infinitivu
 - části slov, spojení slov
- heslo² = heslová stať, entry

Heslář

- rozsah
- výběr podle oboru a typu
- obecný jazyk: frekvence

Heslář

- obecná slova
 - běžná slova (varianty)
 - zkratky
 - části slov
 - víceslovné výrazy
- vlastní jména
 - osoby, místa, metonymie, národnosti/skupiny, organizace, náboženství, předměty
- zkratky vs. plné názvy
- slovní spojení samostatně?

Heslář

- Achilles
- SSJČ: jm. řeckého reka v Homérově Iliadě: Achillova pata, přen. zranitelné místo; každý člověk má svou Achillovu patu; med. Achillova šlacha upínající se na kost patní;
- SSČ: Achillova pata, zranitelné místo; Achillova šlacha, šlacha lýtkového svalu upínající se na patní kost
- všechna slova v definici musejí být v hesláři

Dictionary Writing Systems

- aplikace pro tvorbu slovníků (obvykle celý proces tvorby)
- často vlastní
- komerční
 - IDM DPS klient server (Windows)
 - iLex jádro a dokupované moduly, samostatně nebo klient server, mobility (Windows, Linux, Mac)
 - TLex online, offline (Windows, Mac)
- nekomerční (Glossword, Matapuna)
- DEB (Dictionary Editor and Browser)

iLEX v 2.1.039 Licensed to Jens Erlendsen, Copyright (c) 2004-2009 EMP ApS Username: admin

Server-iLexWeb-EMP

Documents

- Link Up
- Design
- 0 Super Toc (id=050)
- 1 All Tocs (id=052)
- About us (id=020)
- 2.1 Advanced searching (id=021)
- Asset Panel (id=032)
- 2.1 Change tracking (id=008)
- Clipboard Panel (id=033)
- Conforming XML (id=012)
- Cross-references (id=020)
- Dictionaries encyclopedias (id=030)
- Download Presentations (id=035)
- Downloads (id=028)
- Downloads Introduction (id=033)
- Downloads Java (id=066)
- Early customized editing views (id=018)
- element panel (id=043)
- Ergonomic XML editing (id=013)
- Erlendsen Media Publishing (EMP) (id=030)
- Extract Panel (id=047)
- Flexible Design (id=015)
- Full multimedia support (id=017)
- Home (id=010)
- iLEX (id=025)
- iLEX, the background (id=023)
- iLEX, the integrated XML editing system (id=042)
- iLEX an overview (id=042)
- iLEX System Introduction (id=043)**
- 2.1 Improved Graphical Statistics (id=056)
- Integrated databases with separate data (id=018)
- Integrated Lexicography, Editing and XML (id=044)
- 2.1 Legal texts (id=063)
- Lexicography (id=058)
- Library Introduction (id=038)
- Library TOC (id=032)
- List Panel (id=054)
- 2.1 Lists (id=053)
- Metadata separated from content (id=037)
- 2.1 Multiple, customized keyboards (id=009)
- 2.1 Namespace (id=030)
- New iLEX version 2.1 (id=028)
- Powerful database (id=014)
- Pre-cooked search specifications (id=036)
- Reference Panel (id=051)
- 2.1 Schemata (id=005)
- search panel (id=048)
- 2.1 SmartEdit (id=055)
- 2.1 Spell checking (id=007)
- 2.1 Statistics on Schema usage (id=022)
- Statistics Panel (id=045)
- Structured Editing, introduction (id=058)
- Support (id=027)
- Support Introduction (id=034)
- 2.1 Technical documentation (id=061)
- 2.1 TEI - Text Encoding Initiative (id=052)

Integrated Lexicography, Editing and XML

iLexWeb=100000044

The main purpose of a workstation is integration, which for iLEX primarily means integration of the database and the editing system. A second purpose is that various functions need to be applied to data during an editing cycle. Thirdly, there is the ability to work simultaneously, not only with multiple documents from the same database, but also with multiple documents from multiple databases.¶

Integration is supported by the main layout of the user interface. On the left hand side, a Project Panel gives access to all the documents in an open database; and also to the databases themselves if more than one is open. Note that access does not imply write permission.¶

<picture>¶

The project panel has a tab providing access to the design documents for the project. These can be opened and edited just as any other document.¶

The panel for document windows is in the middle. As multiple documents from several projects can be open simultaneously, a model with one document per window has been selected. The view of a document in a window can be changed and moved, as discussed in the following.¶

On the right hand side there is a task panel that provides various tabs for different tools etc., which can be used for the document currently in focus, or for a particular editing task on several documents from the database.¶

Currently the following panels are provided:¶

- Element panel
- Search panel
- Extract panel
- Statistics panel
- Validation panel
- Template panel
- Reference panel
- Asset panel
- Clipboard panel

iLEX System Introduction (id=043)

iLexWeb=100000044

iLEX System Introduction

iLEX is programmed in 100 percent Java on top of several open source Java modules. iLEX itself comprises about 3000 modules, including the database. More than 400 functions and user interface items can be controlled (on/off) in an installation, so the software is comprehensive, not a goal in itself – merely a fact.¶

The database uses Unicode internally. Currently we have no way of mapping between symbolic character entities and the hex codes in the database, but Unicode private use areas can be used, as well as a number of add-ons, e.g. the characterPad, characterMap and characterLab for typing in characters not available from the keyboard.¶

A data model for document windows is in the middle. As multiple documents from several projects can be open simultaneously, a model with one document per window has been selected. The view of a document in a window can be changed and moved, as discussed in the following.¶

On the right hand side there is a task panel that provides various tabs for different tools etc., which can be used for the document currently in focus, or for a particular editing task on several documents from the database.¶

Currently the following panels are provided:¶

- Element panel
- Search panel
- Extract panel
- Statistics panel
- Validation panel
- Template panel
- Reference panel
- Asset panel
- Clipboard panel

Validate panel

Current error

Document: Downloads (id=028)

Error: cvc-complex-type.4

Explanation:

Attribute 'docID' must appear on element 'content_inf'.

Result

Fatal errors: 0 Errors: 2 Warnings: 0

Time: 0:00:06.143 Documents: 2

Document:	Error:	Explanation:
Downloads (id=028)	cvc-complex-type.4	Attribute 'docID'
Downloads Java (id=066)	cvc-complex-type.2.4.8	The content c

Element Statistics View - Server-iLexWeb-EMP

document 55

lib 56

content supadic 56

RPCDATA 56

[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 (*)
sangue (*)
sani
sans (*)
sans-cœur (*)
sans-joie (*)
Santa Claus (*)
santé (*)
saoul
saper [1] (*)
saper [2]
sapré (*)

sans (*)
sans-cœur (*)
sani

Lemma: sans LemmaSign=sans,Modified=2009-02-23 20
-Pronunciation: text 'sɔ'
-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] sɔ
POSGroup:
LemmaSign:
PartOfSpeech: prep.

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 [sɔ] 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'aurais 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 [sɔkœr] 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 [sɔʒwa] n.m.
1 great blue heron <Loc: Lv68, Re31> [Admin]

Santa Claus [sɔtaklɔz, sɛteklɔz] n.prop.
1 Santa Claus <Loc: AC, EV, IB, Lv68, Ph36> [Admin]

santé [sɛte] 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> <Loc: AL, LF, Da84, Lv68> [Admin]

IDM DPS

The screenshot displays the 'Entry Editor' window for 'dps2 server'. The interface includes a menu bar (FILE, EDIT, CONFIG, TOOLS, HELP), a toolbar with icons for NEW, PRINT, SAVE, TEMPLATE, VERSIONS, and FIND, and a status bar showing 'User: idm', 'Project: quick', 'Headword: quick', 'Senses: 5', 'Words: 205', 'Characters: 1110', and '1110'.

The main content area is split into two panes. The left pane shows the dictionary entry for 'quick':

quick ■ **adjective** 1 moving fast or doing something in a short time. ► lasting a short time. ► prompt. 2 intelligent. ► (of one's eye or ear) alert. 3 (of a person's temper) easily roused. ■ **noun** 1 (the quick) the tender flesh below the growing part of a fingernail or toenail. ► the central or most sensitive part of someone or something. 2 [as pl n the quick] archaic those who are living. – PHRASES **cut someone to the quick** cause someone deep distress. **quick and dirty** informal chiefly US done or produced hastily. **a quick one** informal a rapidly consumed alcoholic drink. **quick with child** archaic at a stage of pregnancy when the fetus can be felt to move. – DERIVATIVES **quickly** adverb **quickness** noun – ORIGIN OE *cwīclic*, *cwīclic* alive, animated, alert, of Gmc origin.

The right pane shows the morphological analysis of the word 'quick' using the XREF tool. The analysis tree is as follows:

- E ►
 - HG
 - RW **quick**
 - SG
 - SE1
 - POSG
 - POS ►
 - SE2
 - HSDICT ►
 - DF moving fast or doing something in a short time.
 - EG ►
 - EX some children are particularly quick learners
 - EX I was much quicker than him and held him at bay for several laps
 - SG with SY infinitive /sv
 - EX he was always quick to point out her faults.

The right pane also features an 'ATTRIBUTES' section with a table:

Element	DF
Name	Value
suppressed	- undefined

Below the table are sections for 'SIGN OFF' and 'REVISION FILE'. The 'REVISION FILE' section contains a text area with the text 'CHECK IT TWICE' and buttons for 'CANCEL' and 'OK'.