Vybrané pojmy z oblasti selekčních jazyků a věcného pořádání informací

Materiál určený pro studenty předmětu "Selekční jazyky"

Zpracoval Josef Schwarz

říjen 2003

upraveno září 2006

(text vznikl úpravou výtahu z práce SCHWARZ, J. *Vývoj teorie a praxe tezaurů v České republice : nástin dějin deskriptorových selekčních jazyků v bývalém Československu se zaměřením na vývoj teoretických, metodických a normativních aspektů tvorby tezaurů*. Diplomová práce ÚISK FF UK. Praha : vlastním nákladem, 1999. IX, 121 s.)

Poznámka: tento text slouží výhradně pro studijní potřebu, nelze jej dále šířit ani používat jako pramen při psaní seminárních a dalších prací.

OBSAH

ZÁKLADNÍ POJMY

Oblast věcného pořádání informací a selekčních jazyků patří do širší problematiky pořádání informací a informačních jazyků.

Pořádání informací je jedním z charakteristických procesů informačního systému, který na obecné úrovni zahrnuje procesy výběru a akvizice informací, vstupního zpracování, uložení informací a výstupního zpracování.

Informační systém je souhrn prvků, jejich vztahů a vlastností (obecně složek informačního systému), který jako celek slouží pro získávání, uchovávání a šíření informací. Složky informačního systému můžeme analyzovat na základě obecné teorie systémů a vyčlenit tak ***prvky***, ze kterých se informační systém skládá, a ***procesy***, které probíhají v rámci informačního systému, popř. v rámci interakce informačního systému s jeho okolím. Prvky informačního systému dále můžeme rozložit na subsystémy informačního systému, které jsou jeho funkční součástí, a objekty informačního systému, které jsou předmětem informačního systému. Objekty informačního systému jsou informační objekty, jimiž na konkrétní úrovni rozumíme např. dokumenty.

ZNAK, POJEM, VÝZNAM

Disciplína zkoumající vlastnosti znaků a znakových soustav, které nesou určitý význam, se nazývá **sémiotika**.

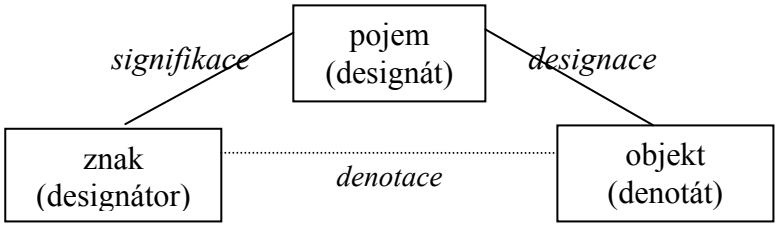
Znak je základní sémiotická jednotka chápaná jako třída smyslově vnímatelných signálů, které poukazují k témuž objektu, vlastnosti nebo stavu, resp. které je na základě konvence zastupují.

Pojem je myšlenková konstrukce vzniklá abstrakcí na základě vlastností společných určité množině objektů; myšlenková představa je určená svou **intenzí** (obsahem), tj. souhrnem podstatných charakteristik množiny denotátů, a **extenzí** (rozsahem), tj. množinou objektů (denotátů), kterou daný pojem zahrnuje.

Denotát je součást (jev, předmět, proces atd., obecně entita) objektivní reality, který je zastoupen znakem a „myšlen“ pojmem.

Základní vztahy mezi objektivní realitou, myšlením a jazykem vyjadřuje model, který se nazývá **sémiotický (sémantický) trojúhelník**.

Obr. č. 1 Sémiotický (sémantický) trojúhelník



Sémiotický trojúhelník vyjadřuje základní představu, že znaky se nevztahují k reálnému objektu přímo, ale prostřednictvím abstraktní představy, pojmu. Místo **pojmu** se v sémiotickém trojúhelníku někdy uvádí **význam či smysl**, místo **znaku** se někdy uvádí obecnější **symbol** nebo konkrétnější **slovo** či **termín**. Pod **objektem** se rozumí **konkrétní věc či předmět nebo abstraktní entita**, označuje se někdy také jako **referent** či **nominát**. Vztah mezi pojmem a objektem je myšlenkový a označuje se jako designace, vztah mezi znakem a pojmem je významový a označuje se jako signifikace, vztah mezi znakem a objektem je označovací a používá se pro něj termínu

denotace.

Významem znaků se zabývá část sémiotiky nazývaná **sémantika**, v lingvistice je tato oblast označována spíše jako **lexikální sémantika**.

Význam je pojmová hodnota, obsah jazykového znaku. **Smysl** je systém významových vztahů; způsob, jímž je předmět označený znakem „podán" (Jitřenka/Večernice - Frege).
Sémantém je nejmenší jazykový znak vyjadřující lexikální význam; často se kryje s kořenem slova. **Sém** je nejmenší jednotka významu, k níž se dospívá sémantickou analýzou; v jazykovém systému pro sém neexistuje odpovídající formální jednotka. **Sémém** je množina sémů; v jazykovém systému je vyjádřen lexémem (lexikální jednotkou).

JAZYK

Jazyk je „systém znaků umožňujících komunikaci, který se obvykle skládá ze slovní zásoby a z pravidel" (ČSN ISO 5127-1/1.1.2-01).

Vzhledem k tomu, že řada procesů a jevů je v rámci informačního systému účelově *formalizována*, používá se v mnoha případech pro komunikaci informací **umělý jazyk**, kterým rozumíme „jazyk vytvořený nebo řízený pomocí souboru předem stanovených pravidel" (ČSN ISO 5127-1/1.1.1-03). Rozdíl oproti **přirozenému jazyku** spočívá v tom, že v přirozeném jazyce slovní zásoba (lexikum) i soubor pravidel (gramatika) nebyly stanoveny předem, ale mají svůj specifický genetický původ, tzn. *přirozený jazyk* je „jazyk, který se vyvíjí a jehož pravidla vyplývají z úzu, takže nemusejí být formálně stanovena" (ČSN ISO 5127-1/1.1.1-02).

Přestože mezi umělým (v našem případě selekčním) a přirozeným jazykem existuje řada rozdílů a někteří autoři dokonce umělý jazyk nepovažují za jazyk ve vlastním slova smyslu,[1] lze pro popis umělého (selekčního) jazyka použít lingvistickou terminologii, i když lingvistika samotná považuje za hlavní předmět svého zájmu přirozený jazyk.[2]

Lexikum je slovní zásoba určitého jazyka. Základní jednotkou lexika je **lexikální jednotka**, jež může být definována jako minimální posloupnost fonémů nebo grafémů, která je dále sémanticky nedělitelná. Pokud se lexikální jednotka skládá z více než jednoho slova, označuje se jako *sousloví.* **Sousloví** je lexikální jednotka, kterou lze morfologicky rozčlenit na oddělené složky; sousloví se skládá ze základu, kterým je obvykle substantivum, a modifikátoru, kterým bývá adjektivum, neshodný přívlastek nebo jiný prvek.

V oblasti selekčních jazyků za základní jednotku slovníku považujeme **pořádací znak**. Pořádací znak systematického selekčního jazyka se nazývá **klasifikační znak**; klasifikační znak je vyjádřen **notací**, tj. posloupností znaků, který se používá k označení tříd klasifikačního schématu (**třída** je skupina klasifikovaných pojmů vzniklá na základě shodných charakteristik za účelem definování sémantických vztahů mezi nimi; třída je označena notací. Tvoří základní prvek klasifikačního systému označující základní kategorii.) Podle druhu použitých znaků dělíme notaci na **alfabetickou** (jako znaky notace jsou použita písmena), **numerickou** (jako znaky jsou použity číslice) nebo **alfanumerickou** (smíšenou – jsou použita písmena i číslice). Podle struktury rozlišujeme **notaci expanzivní** (tj. notaci umožňující rozšiřování klasifikačního systému), lineární notaci (vyjadřuje pořadí tříd, nikoliv však vztahy mezi nimi) a **hierarchickou notaci** (vyjadřuje vztahy mezi třídami). **Desetinná notace** je typ numerické notace užívající číslic 0-9 a umožňující

[1] „Označení 'jazyk' přísluší pouze přirozenému lidskému dorozumívacímu kódu." V ostatních případech „by se mělo raději mluvit o 'umělých jazykových kódech' nebo podobně" (ERHART, A. *Základy jazykovědy*. Praha, Státní pedagogické nakladatelství, 1984, s. 12.).

[2] „Centrem zájmu jazykovědců je však pochopitelně v první řadě **přirozený jazyk** jako základní, nejbohatší a polyfunkční prostředek komunikace" (ČERMÁK, F. *Jazyk a jazykověda : přehled a slovníky.* Praha : Karolinum, 2001, s. 15.).

logický rozklad předmětu dokumentu na jednotlivé komponenty. Každé číslo notace je chápáno jako desetinný zlomek s vypuštěnou desetinnou čárkou.

Pořádací znak předmětového selekčního jazyka se nazývá **lexikální jednotka**. Lexikální jednotky deskriptorového selekčního jazyka (tedy lexikální jednotky v tezauru) jsou dvou typů: **deskriptor** (resp. *nepreferovaný termín*[3]) je „lexikální jednotka užívaná závazně při indexování k vyjádření určitého pojmu“ (ČSN 01 0193, 1996:5), **nedeskriptor** (resp. *nepreferovaný termín*[4]) je „ekvivalent nebo kvaziekvivalent preferovaného termínu; nepreferovaný termín není dokumentu přiřazován, ale slouží jako uživatelský vstup do tezauru nebo abecedního rejstříku; uživatel je odkázán příslušným pokynem (např. viz) k ekvivalentnímu preferovanému termínu“ (ČSN 01 0193, 1996:5). Lexikální jednotka předmětového selekčního jazyka založeného na předmětových heslech se nazývá **předmětové heslo**; skládá se z hesla, podhesla a doplňku.

Gramatika je soubor pravidel určujících způsob tvorby tvarů slov a jejich spojování do vět. **Syntax** je část gramatiky, která se zabývá skladbou vět a souvětí. **Morfologie** je nauka o druzích slov, o jejich tvarech a o významech tvarů.

Gramatika selekčního jazyka je soubor pravidel a prostředků, kterými se tvoří různé tvary pořádacích znaků a jimiž se řídí jejich spojování do vyšších jednotek při indexaci či klasifikaci. **Syntax selekčního jazyka** je část gramatiky selekčního jazyka, která určuje pravidla pro spojování pořádacích znaků. **Morfologie selekčního jazyka** je část gramatiky selekčního jazyka, která určuje pravidla pro modifikaci tvarů pořádacích znaků pomocí specifických morfologických prostředků.[5]

Homonymie je „vlastnost dvou nebo více termínů, které mají stejnou grafickou nebo zvukovou podobu, ale rozdílný význam“ (ČSN ISO 5127-1/1.1.2-16). Speciálním případem homonymie je **homografie**, kterou se rozumí „vlastnost dvou nebo více termínů, které mají stejnou grafickou formu, ale rozdílný význam“ (ČSN ISO 5127-1/1.1.2-16). **Polysémie** je „vlastnost slova, které má dva nebo více etymologicky příbuzných významů“ (ČSN ISO 5127-1/1.1.2-14). Pro selekční jazyky je nejzávažnějším jevem homografie, přičemž z hlediska řešení mnohovýznamovosti v rámci slovníku selekčního jazyka je bezpředmětné rozlišení mezi homografií a polysémií.

Synonymie je „vlastnost dvou nebo více termínů majících odlišnou formu a přesně nebo přibližně tentýž význam“ (ČSN ISO 5127-1/1.1.2-12).

Hierarchický vztah je „formální vztah mezi dvěma termíny nebo třídami, kde jeden (jedna) je podřízen (podřízena) druhému (druhé)“ (ČSN ISO 5127-6/3.4.4-09). **Asociativní vztah** je „sémantický vztah mezi pojmy se vzájemnou vazbou z hlediska specifického účelu“ (ČSN ISO 5127-6/3.4.4-04). **Vztah ekvivalence** je „formální vztah mezi termíny představovanými stejným (stejnými) deskriptorem (deskriptory) nebo znakem třídy“ (ČSN ISO 5127-6/3.4.4-11).

PROCESY VĚCNÉHO POŘÁDÁNÍ INFORMACÍ

Pořádání informací je dílčím procesem vstupního zpracování informací, jenž probíhá

[3] Norma upřednostňuje termín preferovaný termín, nicméně z praktických důvodů se používá termínu deskriptor.

[4] Norma upřednostňuje termín nepreferovaný termín, nicméně z praktických důvodů se používá termínu nedeskriptor.

[5] Zejména v tezauru se používá těchto gramatických prostředků: spoje (na syntaktické úrovni) nebo role a váhy (na morfologické úrovni); jsou vyjádřeny ve formě indikátorů. **Spoj** je „znak nebo symbol použitý ke spojení deskriptorů přiřazených dokumentu nebo rešeršnímu požadavku a zabraňující náhodnému spojení těchto deskriptorů s jinými“ (ČSN ISO 5127-3a/3.3.2-06). **Indikátor role** je „pomocný symbol, který může být vybrán ze zvláštního seznamu a připojen k deskriptoru pro vyjádření, ve kterém smyslu byl deskriptor použit“ (ČSN ISO 5127-6/3.4.2-11). **Indikátor váhy** je pomocný symbol, který na základě určité škály vyjadřuje důležitost deskriptoru z hlediska obsahu dokumentu a v souvislosti s dalšími lexikálními jednotkami selekčního obrazu dokumentu.

v rámci informačního systému. Procesem předcházejícím pořádání informací je informační analýza dokumentů, kterou se zjišťují významné identifikační a obsahové charakteristiky dokumentu. Pořádání informací je vytváření organizovaných souborů informací na základě určitého systému. Pořádání informací lze rozdělit na **identifikační pořádání informací**, při kterém jsou zjišťovány formální charakteristiky dokumentu, a **věcné pořádání informací**, při kterém jsou zjišťovány obsahové charakteristiky dokumentu.

Věcné pořádání informací lze podle základní charakteristiky použitého systému pořádání rozdělit na systematické pořádání informací a předmětové pořádání informací. **Systematické pořádání informací** je proces, při kterém jsou informace vřazovány na dané místo v rámci systematicky uspořádaného souboru (systému) lidského poznání, přičemž jejich postavení se v zásadě řídí rodo-druhovými vztahy a slovní formulace obsahu dokumentu bývá většinou nahrazena znaky umělého jazyka (notacemi)[6]. Pro systematické pořádání informací se používá také označení třídění, klasifikace, systematické zpracování, systematická katalogizace ad.

Předmětové pořádání informací je proces, při kterém jsou informace vyjádřeny souborem abecedně uspořádaných hesel[7]. Pro předmětové pořádání informací se používá také označení heslování, předmětové třídění, předmětová klasifikace, předmětové zpracování ad. Používání termínů označujících předmětové pořádání informací jako třídění, resp. klasifikaci není vhodné, protože způsobuje terminologickou konfúzi.

Indexace (indexování, *indexing*) je „proces vyjádření výsledku analýzy dokumentu prostřednictvím prvků selekčního jazyka nebo přirozeného jazyka, obvykle s cílem umožnit zpětné vyhledávání“ (ČSN ISO 5127-3a/3.2.1-03). **Automatická indexace** (automatické indexování, *automated indexing*) je „vyjádření obsahu dokumentu pomocí automatického výběru slov nebo termínů z textu nebo pomocí automatického přiřazování termínů selekčního jazyka“ (ČSN ISO 5127-3a/3.3.3-01). **Poloautomatická indexace** (*machine-aided indexing*) je ekvivalentně k předchozí definici vyjádření obsahu dokumentu pomocí poloautomatického výběru slov nebo termínů z textu nebo pomocí poloautomatického přiřazování termínů selekčního jazyka, přičemž poloautomatickými postupy rozumíme takové procedury, při kterých je část procesu indexace provedena automaticky a výsledek této části slouží jako podklad pro intelektuální indexaci.

Postkoordinovaná indexace je indexace bez předem stanoveného uspořádání pořádacích znaků (lexikálních jednotek nebo klasifikačních znaků). Ke koordinaci (kombinaci) pořádacích znaků dochází až při vyhledávání. **Prekoordinovaná indexace** je indexace dokumentů, při které je uspořádání pořádacích znaků dáno selekčním jazykem.

Klasifikace je přidělování notací (znaků tříd) klasifikačního systému za účelem vyjádření obsahu dokumentu.

Fazetace je rozdělení slovníku selekčního jazyka pomocí faset. **Fazeta** je kategorie entit vytvořená uplatněním jedné klasifikační charakteristiky (principium divisionis), která je pro danou kategorii (třídu) podstatná, strukturální. Fazety vyjadřují vlastnosti použité pro seskupování pojmů podle jejich podstaty. Zjednodušeně lze říci, že **fazeta** je velmi obecná kategorie, která se používá pro rozdělení slovníku selekčního jazyka podle základních charakteristik pojmů, jenž jsou vyjádřeny konkrétními pořádacími znaky.

TYPOLOGIE SELEKČNÍCH JAZYKŮ

Umělý jazyk, používaný v rámci informačního systému, se nazývá **informační jazyk**. Podle funkce informačního jazyka lze vydělit **algoritmické informační jazyky**, kterými rozumíme

---

[6] volně podle: KOVÁŘ, B. *Věcné pořádání informací a selekční jazyky. Díl 1. Úvod do problematiky, systematické pořádání.* Praha : ÚVTEI, 1981, s. 9-10.

[7] volně podle: KOVÁŘ, B. *Věcné pořádání informací a selekční jazyky. Díl 2. Předmětová pořádání, mezinárodní spolupráce, automatické indexování.* Praha : ÚVTEI, 1982, s 5.

programovací jazyky, **logické informační jazyky**, které se používají k formalizaci pojmů pomocí matematické logiky (např. dotazovací jazyky), a **selekční informační jazyky**, používané pro zaznamenávání, třídění, ukládání a vyhledávání informací. Selekční informační jazyky lze rozdělit na **identifikační selekční informační jazyky**, které slouží pro popis formálních charakteristik dokumentu, a **věcné selekční informační jazyky**, které slouží pro popis obsahových charakteristik dokumentu. Identifikačními selekčními informačními jazyky se nebudeme dále zabývat. Věcné selekční informační jazyky budeme dále označovat jako *selekční jazyky*.

Obecná definice **selekčního jazyka** jej charakterizuje jako „formalizovaný jazyk používaný k charakterizování dat nebo obsahu dokumentů za účelem jejich ukládání a vyhledávání“ (ČSN ISO 5127-6/3.4.1-01). Cizojazyčné ekvivalenty pro pojem selekčního jazyka jsou *information retrieval languages* (angličtina), *informacionno-poiskovyj jazyk* (ruština), *Informationsrecherchesprache* (němčina), *langage de recherche documentaire* (francouzština). Tato terminologie se ovšem v zahraniční ani v československé literatuře nepoužívala jednotně a pro označení pojmu selekčního jazyka lze nalézt další výrazy jako např. průzkumový jazyk, informačně-selekční jazyk, informační selekční jazyk, informační jazyk, dokumentační jazyk, dokumentační selekční jazyk, rešeršní jazyk, informačně-vyhledávací jazyk, bibliografický jazyk, katalogizační jazyk nebo systémy pořádání. Dále důsledně používáme termínu selekční jazyk.

Vzhledem k existenci různých selekčních jazyků je žádoucí rozdělit je podle druhů a typů. Typologie selekčního jazyka však není jednotná, protože nelze stanovit jednoznačné *principium divisionis* pro rozdělení jednotlivých selekčních jazyků. Selekční jazyky můžeme rozdělit podle jejich funkce, vnitřní struktury, uspořádání pojmů, stupně formalizace, šířky tematického zaměření, expanzivity a dalších kritérií. V teorii i praxi selekčních jazyků se nejvíce osvědčilo rozdělení podle toho, jak jsou v selekčním jazyce uspořádány jednotlivé pojmy, a podle toho, jakým způsobem jsou vytvářeny selekční obrazy dokumentů (**selekční obraz dokumentu** je množina všech pořádacích znaků přiřazených dokumentu).

Na základě těchto principů můžeme rozdělit selekční jazyky na dva základní druhy: *systematické selekční jazyky* a *předmětové selekční jazyky*, a dva základní typy: *prekoordinované selekční jazyky* a *postkoordinované selekční jazyky*.

Systematický selekční jazyk je selekční jazyk používaný „pro strukturní zpracování dokumentů nebo dat pomocí symbolů a příslušných termínů s cílem umožnit systematický přístup, v případě potřeby s pomocí abecedního rejstříku“ (ČSN ISO 5127-6/3.4.1-03).[8] Pro tento pojem se také používají označení klasifikační systém, knihovnicko-bibliografická klasifikace, bibliografický klasifikační systém, klasifikace (ve smyslu systému, nikoliv procesu), třídění (ve smyslu systému, nikoliv procesu), knihovnické třídění, systematické třídění, systematická pořádací soustava, ad.

Předmětový selekční jazyk je (ekvivalentně[9] definici systematického selekčního jazyka) selekční jazyk používaný pro strukturní zpracování dokumentů nebo dat pomocí abecedně uspořádaných termínů s cílem umožnit předmětový přístup. Pro tento pojem se také používají označení systémy heslování, systémy předmětových hesel, předmětové třídění, abecední předmětové třídění, předmětová pořádací soustava, předmětová klasifikace ad. Používání termínů označujících předmětový selekční jazyk jako třídění, resp. klasifikaci není vhodné, protože způsobují terminologickou konfúzi.

Prekoordinovaný selekční jazyk je selekční jazyk, jehož lexikum se skládá[10] z *pořádacích znaků,* které vyjadřují složené pojmy a které jsou používány pro indexaci i vyhledávání.

Postkoordinovaný selekční jazyk je selekční jazyk, jehož lexikum se skládá[11] z *pořádacích znaků*, jež vyjadřují jednoduché pojmy, při indexaci jsou do selekčního obrazu dokumentu zařazovány nezávisle na sobě a k jejich kombinaci dochází až v průběhu vyhledávání.

---

[8] Norma ovšem neuvádí termín *systematický selekční jazyk,* ale *klasifikační systém.*
[9] Definici předmětového selekčního jazyka norma ČSN ISO 5127-6 neobsahuje.
[10] Nikoliv nutně, ale charakteristicky.
[11] Nikoliv nutně, ale charakteristicky.

Na základě výše uvedené typologie a definic můžeme odvodit označení a význam pro čtyři základní kategorie selekčních jazyků: *prekoordinovaný systematický selekční jazyk, postkoordinovaný systematický selekční jazyk, prekoordinovaný předmětový selekční jazyk* a *postkoordinovaný předmětový selekční jazyk.* K prekoordinovaným systematický selekčním jazykům[12] patří např. Deweyho desetinné třídění (DDT) nebo Mezinárodní desetinné třídění (MDT), mezi postkoordinované systematické selekční jazyky[13] se řadí např. Ranghanatanovo dvojtečkové třídění.

Předmětové selekční jazyky dále můžeme rozdělit na tři dílčí typy podle charakteru lexikálních jednotek.

První skupinou předmětových selekčních jazyků jsou **předmětové selekční jazyky založené na použití slov z názvu dokumentů**. Podle toho, zda vznikly intelektuálním nebo automatizovaným zpracováním, se rozlišují názvové katalogy, resp. názvové rejstříky, a permutované (cyklické) rejstříky dvou typů: **KWIC** (Keyword in Context) a **KWOC** (Keyword out of Context). Tento typ předmětového selekčního jazyka se označuje někdy také termínem **klíčová slova**, který ovšem může v různých kontextech nabývat různých významů,[14] proto je vhodnější použít přesnějšího označení *klíčová slova z názvů dokumentů.*

Druhým typem předmětového selekčního jazyka je *předmětový selekční jazyk typu předmětových hesel*, pro jednoduchost označovaný často jako **předmětová hesla**. Lexikum tohoto typu selekčního jazyka sestává z předmětových hesel. Protože je jednoduchý pořádací znak tohoto selekčního jazyka, předmětové heslo, strukturován na dílčí syntakticky spojené složky (heslo, podheslo, doplněk), jedná se o prekoordinovaný selekční jazyk.

Třetí skupinou předmětových selekčních jazyků jsou *předmětové selekční jazyky deskriptorového typu,* pro jednoduchost označované často termínem **deskriptorové selekční jazyky**. Lexikum těchto selekčních jazyků je tvořeno lexikálními jednotkami, jejichž struktura a význam jsou specificky vymezeny a jenž určují i charakter deskriptorového selekčního jazyka jako postkoordinovaného selekčního jazyka. Můžeme vyčlenit dva základní dílčí typy deskriptorových selekčních jazyků, deskriptorové selekční jazyky založené na unitermech a deskriptorové selekční jazyky založené na deskriptorech.

Jak jsme uvedli výše, selekční jazyky můžeme rozdělit do dalších kategorií podle nejrůznějších kritérií.

Podle šířky tematického zaměření rozlišujeme **univerzální selekční jazyky**, jejichž lexikum zahrnuje celé univerzum lidského poznání, a **speciální selekční jazyky**, jejichž lexikum zahrnuje určitou, většinou oborově vymezenou oblast lidského poznání. Speciálním typem univerzálních selekčních jazyků jsou **polytematické**, resp. **polytechnické selekční jazyky**, zaměřené na vybranou oblast lidského poznání zahrnující několik vymezených oborů (v případě polytechnických selekčních jazyků se jedná o obory technické). V oblasti speciálních selekčních jazyků lze vydělit **oborové**, resp. **odvětvové selekční jazyky**, tzn. selekční jazyky zahrnující lexiku vybraného oboru nebo odvětví národního hospodářství. Zde je nutno podotknout, že princip univerzálnosti je charakteristický spíše pro systematické selekční jazyky, kdežto princip speciálnosti se uplatňuje především v předmětových selekčních jazycích.

Podle toho, zda jsou v selekčním jazyce uplatněny gramatické prostředky, rozlišujeme **selekční jazyky s gramatikou** a **selekční jazyky bez gramatiky**, resp. *selekční jazyky s nulovou gramatikou*.

---

[12] Z hlediska vnitřního uspořádání lexika se jedná o *systematické selekční jazyky hierarchického typu*, označované někdy méně přesně jako *hierarchické klasifikace*.

[13] Z hlediska vnitřního uspořádání lexika se jedná o *systematické selekční jazyky fazetového typu*, označované také někdy jako *fazetové klasifikace*.

[14] Klíčovými slovy se rozumí např. výrazy vybrané z textu dokumentu apod.

TYPY DESKRIPTOROVÝCH SELEKČNÍCH JAZYKŮ A TEZAURŮ

Jak jsme uvedli výše, dva základní typy deskriptorových selekčních jazyků jsou deskriptorové selekční jazyky založené na unitermech, zkráceně *unitermy*, a deskriptorové selekční jazyky založené na deskriptorech, které se v praxi běžně označují jako tezaury – to je však chybné, protože tezaurus je pouze jednou složkou deskriptorového selekčního jazyka, a to jeho slovníkem.

Vývojově starším, tezaurům předcházejícím systémem, jsou unitermy. **Uniterm** je „nejmenší významový prvek selekčního jazyka použitý k vyjádření specifického pojmu v rámci systému koordinovaného indexování“ (ČSN ISO 5127-6/3.4.2-10). Systém unitermů je charakterizován lexikem, jehož lexikální jednotky (unitermy) jsou vyjádřeny většinou jednoslovně, v nezbytných případech[15] souslovím, jenž může obsahovat vzájemné vztahy lexikálních jednotek, které odstraňují synonymii a homonymii, které však nezahrnuje hierarchické vztahy lexikálních jednotek. Všechny unitermy jsou považovány za lexikální jednotky se stejnou hierarchickou úrovní. Systém unitermů prošel od svého vzniku v r. 1951 v průběhu 50. let určitým vývojem a uvedený popis systému odpovídá jeho konečnému stavu na přelomu 50. a 60. let, kdy byl jeho vývoj v zahraničí uzavřen nástupem tezaurů. V původní verzi např. musely být všechny unitermy jednoslovné a mezi lexikálními jednotkami neexistovaly žádné vztahy.

Z unitermů a dalších systémů se vyvinuly deskriptorové selekční jazyky založené na deskriptorech, jejichž forma a vztahy jsou standardizovány slovníkem se speciální strukturou, tezaurem. **Tezaurus** je „slovník řízeného selekčního jazyka uspořádaný tak, že explicitně zachycuje apriorní vztahy mezi pojmy“ (ČSN 01 0193, 1996:5). Současné pojetí tezauru je zakotveno normou ISO 2788 (1986) doplněnou ISO 5964 (1985) pro vícejazyčné tezaury. Obě normy jsou v českém národním prostředí implementovány jako ČSN 01 0193 (1996) a ČSN 01 0172 (1992).

Podle šířky tematického zaměření rozeznáváme *univerzální tezaury* a *speciální tezaury*. **Univerzální tezaury** zahrnují celé univerzum lidského poznání. Speciálním typem univerzálních tezaurů jsou **polytematické**, resp. **polytechnické tezaury**, zaměřené na vybranou oblast lidského poznání zahrnující několik vymezených oborů (v případě polytechnických tezaurů se jedná o obory technické). Z hlediska struktury i funkce je značně specifickým typem univerzálního tezauru **makrotezaurus**, kterým se rozumí „tezaurus tvořený termíny vysoké úrovně obecnosti a zahrnující širokou oblast (lidského) poznání“ (ČSN ISO 5127-6/3.4.5.1-05). Skutečně univerzální tezaury je velmi obtížné realizovat; praktické pokusy o jejich tvorbu většinou nepřinesly adekvátní výsledky u nás ani v zahraničí.

Tezaury jsou většinou realizovány jako **oborové tezaury**, tzn. tezaury omezené na jeden obor lidského poznání, které jsou nejcharakterističtějšími typy tezaurů.

Podle jazykového zaměření dělíme tezaury na *jednojazyčné tezaury* a *vícejazyčné tezaury*. **Jednojazyčný tezaurus** je tezaurus „obsahující deskriptory a obvykle nedeskriptory převzaté z jednoho přirozeného jazyka“ (ČSN ISO 5127-6/3.4.6.1-01). **Vícejazyčný tezaurus** je tezaurus **„**obsahující deskriptory a obvykle nedeskriptory převzaté z několika přirozených jazyků a vyjadřující ekvivalentní pojmy v každém z těchto jazyků“ (ČSN ISO 5127-6/3.4.6.1-02). Vícejazyčné tezaury jsou někdy označovány nesprávným termínem mnohojazyčné tezaury.[16]

Podle způsobu tvorby, resp. vnitřního uspořádání, vyčleňujeme specifický typ tezauru označovaný termínem **fazetový tezaurus**, kterým rozumíme tezaurus, ve kterém „se vztahy mezi termíny vytvářejí potom, když byly přeskupeny podle faset“ (ČSN ISO 5127-6/3.4.6.1-09). Fazetový přístup může ovlivnit nejen postup tvorby tezauru, ale i uspořádání tezauru.

Z hlediska funkce tezauru rozlišujeme „**tradiční“ tezaurus** (*conventional thesaurus,* resp. „*classic“ thesaurus* nebo „*traditional“ thesaurus*), který se používá pro indexaci a vyhledávání

[15] Jedná se o případy, ve kterých by došlo v důsledku syntaktického rozkladu ke ztrátě významu původního sousloví. Původní verze unitermů však tuto charakteristiku neobsahovala (viz další text).

[16] Uvedený termín je nesprávný proto, že jazyků ve vícejazyčném tezauru nemusí být „mnoho“, ale např. pouze dva.

dokumentů, **vyhledávací tezaurus** (*searching thesaurus, search-aid thesaurus, „advice-giving" thesaurus*), který se používá pouze pro vyhledávání dokumentů,[17] a **indexační thesaurus** (*indexing thesaurus*), který se používá pouze pro indexaci dokumentů[18,19]. Indexační tezaurus se používá v praxi pouze zřídka, větší pozornost se věnuje vyhledávacím tezaurům v souvislosti s možnostmi zpracování plných textů. Vyhledávací tezaurus může být realizován na třech úrovních jako intelektuální, poloautomatická nebo automatická podpora uživatele při sestavování dotazu. Intelektuální podpora pomocí vyhledávacího tezauru spočívá v možnosti intelektuálního výběru lexikálních jednotek z tezauru při sestavování dotazu.[20] Poloautomatická podpora je realizována na základě automatického doplnění dotazu uživatele o automaticky vybrané lexikální jednotky tezauru. Po sestavení dotazu je uživateli automaticky nabídnut seznam potenciálně vhodných lexikálních jednotek (popř. jsou tyto lexikální jednotky automaticky doplněny do dotazu), které souvisejí s již zadanými termíny, s možností jejich potvrzení nebo zamítnutí uživatelem a následným začleněním do dotazu. Automatická podpora může být charakterizována jako analýza dotazu (v přirozeném jazyce) pomocí vyhledávacího tezauru. Po sestavení dotazu (v přirozeném jazyce) se automaticky provede převod z klíčových slov na lexikální jednotky vyhledávacího tezauru a provede se vyhledávání, přičemž uživatel se při práci se systémem ani nemusí dozvědět, že pracuje s tezaurem.

[17] Indexace je provedena jiným tezaurem, jiným selekčním jazykem nebo se jedná o plné texty dokumentů.

[18] Vyhledávání probíhá na základě jiného systému, např. na základě zpracování dotazu v přirozeném jazyce.

[19] AITCHISON, J., GILCHRIST, A., BAWDEN, D. *Thesaurus construction and use : a practical manual. Third ed.* London : Aslib, 1997, s 1-2.

[20] Tato funkce je zcela identická s tradičním tezaurem, rozdíl spočívá v tom, že vyhledávácím tezaurem nejsou indexovány vyhledávané dokumenty.