Řešení staví na následujícím předpokladu, zmíněném v zadání:

> Pokud se jméno se slovesem může kombinovat, vidíme toto spojení v datech.

a na této myšlence:

> Jména označující obecnější koncepty se mohou kombinovat s více slovesy než jména označující konkrétnější koncepty (čili než jména jim podřazená). Proč? Protože kdykoli lze provést děj na specifičtějším konceptu, lze to v jazyku vyjádřit spojením příslušného slovesa buď se jménem označujícím přímo ten specifičtější koncept, ale i se jménem označujícím koncept obecnější. Situaci, kdy například Říša přinese domů pudla, tak můžeme označit slovy „(Říša) přinesl psa“ nebo „(Říša) přinesl (nějaké chundelaté) zvíře“, ne ovšem „(Říša) přinesl bernardýna“. Čili lze-li sloveso použít se jménem J, lze použít s libovolným jménem nadřazeným J, ne však nutně s libovolným jménem podřazeným J. Tolik hypotéza Θ_1.

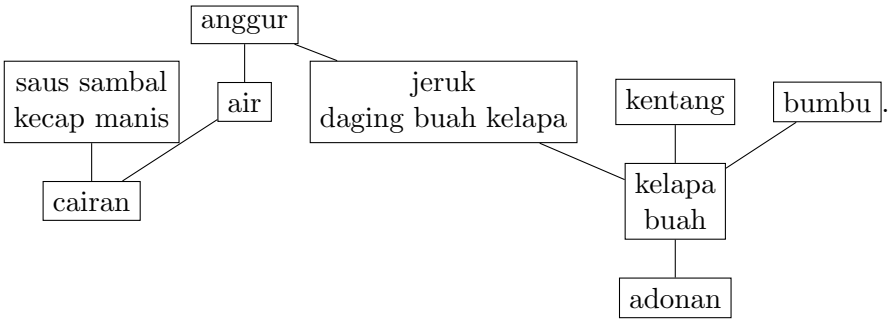
Nyní by tedy mělo stačit sestavit si tabulku souvýskytů pro všechna slovesa a jména, která známe (že slovesa jsou ve spojeních první a jména následují, šlo zjistit ze zadání úkolu 3, podle nějž u dvou slov víme, co jsou zač):

	air	anggur	saus sambal	adonan	jeruk	kentang	kelapa	buah	kecap manis	cairan	daging buah kelapa	bumbu
aduk	0	0	0	1	0	0	0	0	0	0	0	0
bumbui dengan	0	0	1	0	0	0	0	0	1	0	0	0
didihkan	1	1	0	0	0	0	0	0	0	0	0	0
makan	0	1	0	0	1	1	1	1	0	0	1	1
memeras	0	1	0	0	1	0	0	0	0	0	1	0
menghaluskan	0	0	0	0	0	1	0	0	0	0	0	1
minum	1	1	0	0	0	0	0	0	0	0	0	0
potong	0	1	0	1	1	1	1	1	0	0	1	1
sisihkan	0	0	0	1	0	0	0	0	0	0	0	0
tambah	1	1	1	1	1	1	1	1	1	1	1	1
tuang	1	1	1	0	0	0	0	0	1	1	0	0

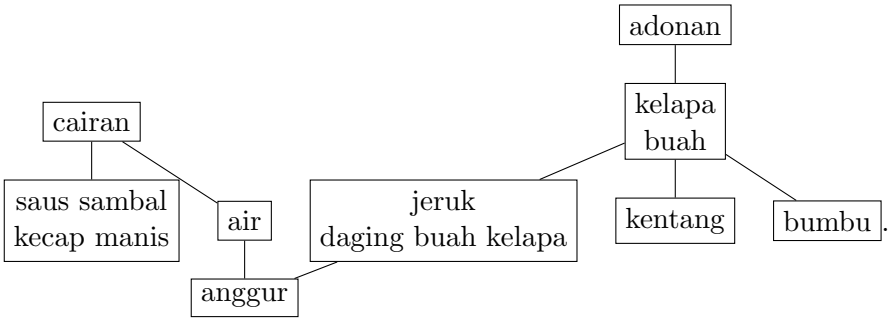
Z tabulky můžeme hned vyčíst dvojice konceptů, které jsou jeden nadřazený druhému (kdykoli jeden obsahuje jedničky všude tam, kde je obsahoval druhý, považujeme jej podle hypotézy Θ_1 za nadřazený):

anggur > air > cairan
{saus sambal, kecap manis} > cairan
{jeruk, daging buah kelapa} > {kelapa, buah}
{kentang, bumbu} > {kelapa, buah}
anggur > {jeruk, daging buah kelapa}

Tyto vztahy můžeme také zakreslit pomocí (tzv. Hasseho) diagramu:



Vida, z diagramu to vypadá, že slova mají často mnoho nadřazených slov (např. *saus sambal*, *kecap manis*, *air* jsou všechny nadřazené slovu *cairan*). To není normální. Hypotéza Θ_1 je tedy špatně; daleko víc by to dávalo smysl, kdyby byl diagram vzhůru nohama:



Teď jsou mnohem přirozeněji *saus sambal*, *kecap manis* a *air* druhy *cairan*. Jak ovšem musíme změnit hypotézu, aby dávala smysl i ona? Zřejmě měla od začátku být takto (Jak snadné je přizpůsobit hypotézu datům!):

> Jména označující specifičtější koncepty mají více vlastností, čili jde s nimi provádět více věcí. Proto se budou pojit s více slovesy než jim nadřazená jména. Například lze „nakrmit psa“, „nakrmit pudla“, ale obvykle lze pouze „přinést pudla“, protože pes obecně může být příliš těžký, než aby ho šlo nosit (jako třeba bernardýn). Navíc každý druh pudla půjde přinést, protože pudly

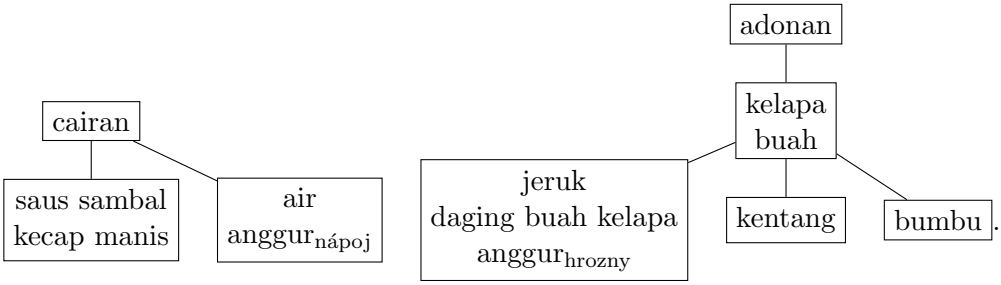
> lze nosit. Protože lidi nazývají obvykle věci jejich jménem a ne příliš obecně (ne třeba „Říša přinesl živočicha“), měla by data doložit, že lze-li sloveso použít se jménem J, lze je použít i s libovolným jménem podřazeným J, ne však nutně s libovolným jménem nadřazeným J.
>
> Tato hypotéza (Θ_2) je opravdu už ta správná, podle které se mělo postupovat. Autor úložky se při psaní autorského řešení spletl, ale rozhodl se původní, mylnou hypotézu ponechat i v autorském řešení. Řešení mu tak přijde informativnější – je v něm vidět, kde se lze snadno splést, a také, jakou úvahou se z toho dostat.

Ovšem ani po převrácení uspořádání a hypotézy všechno nesedí. Zvláště slovo *anggur*, na něž se ptá další otázka, vypadá podezřele. Co za ním vězí?

Víme, že *potong* znamená krájet, a to se vyskytuje právě u všech slov v pravé části hierarchie. Všechny věci napravo jsou tedy krájitelné, ty nalevo nejsou. Navíc nám mohou něco připomínat výrazy „saus sambal“ a „kecap manis“. První zní jako pojmenování pro omáčku (sauce, Sauce, sos), to druhé jako pojmenování pro kečup. *Kecap manis* se možná dá dokonce zahlédnout i v českých obchodech na regálech s asijskými výrobky, a je to opravdu něco jako kečup. Podoba *air* s anglickým výrazem pro vzduch je pravděpodobně náhodná (slovo je krátké), a že *bumbu* má něco s bumbáním, také nezní důvěryhodně. Nejjednodušší vysvětlení, co tedy jsou kategorie nalevo a napravo, je asi to, že jedny jsou tekuté a druhé pevné.

Nyní hledáme nějakou věc (zřejmě ingredienci, když slova pocházejí z kuchařek), která může být tekutá i pevná. (To můžeme považovat za dva odlišné koncepty, tedy tu věc za dvě různé věci. Potom by *anggur* bylo synonymum. Jinak by to musela být ingredience zhruba polotekutá.) Když navíc víme, že *anggur* může *fermentasi*, což zní nápadně podobně slovu *fermentovat*, česky kvasit, už stačí projít si pár známých věcí, které mohou kvasit, a máme odpověď. Je to víno. Vida – i v češtině je to synonymum, zcela přirozeně.

Můžeme si nyní v datech oddělit výskyty slova *anggur* v jeho dvou významech. Dostaneme z toho čistou hierarchii, kde každé slovo má pouze jedno hyperonymum:



Měli bychom nyní ještě oddělit v datech výskyty, které tam byly kvůli jednomu významu slova víno, a ty, které tam byly kvůli tomu druhému. Mohlo by nám to porozházet hierarchii. Naštěstí se ale víno krásně rozdělí mezi dva významy – jeden se stejnou distribucí jako *air* a druhý podobný *jeruk*.

Co se týče filozofičtějších otázek, zodpověděl bych je asi takto:

Slovesa narozdíl od podstatných jmen nemají zdaleka tak členitou hierarchii. Kromě toho, ačkoli to není tak důležitý argument, nemají tak jasně oddělené významy (je jasné, jak se liší jablko od pomeranče, ale třeba slovesa házet, metat, vrhat…se dají používat celkem záměnně). I když v našem malém příkladu jsme jasně uměli postavit jména do hierarchie, čímž jsme mimochodem stejně tak sestavili do hierarchie i přidružená slovesa, s přibývajícím počtem různých jmen/sloves bychom zjišťovali, že mezi slovesy vlastně tak jasná hierarchie vůbec není.

Tím plynule přecházím k druhé myšlence – co by se stalo při aplikování metody na velké množství textů? Jak ze zkušenosti víme, jazykové jednotky nepředstavují vůbec jednoznačné přiřazení slov k významům. Ba naopak, jazyk překypuje synonymií i homonymií, vztah slovo-význam je hodně propletený. Kdybychom uvolnili doménu, ze které budeme spojení vybírat, data by se nám zaplavila synonymy slov, která byla v malé doméně jednoznačná, a jejich homonymy – druhými názvy konceptů, které jsme při zachování malého slovníčku také uměli pojmenovat pouze jediným slovem. Tím by metoda prakticky zkrachovala.

K poslední otázce: metoda by šla zobecnit tak, že by se místo ostrého počítání výskytů v rozmezí {vyskytlo se, nevyskytlo se} počítaly doslova: 1, 2, 3…Ve vzniklém diagramu by pak nebyly hrany vs. nehrany, ale hrany různé váhy. Pod-/nadřazenost by se pak také určovala pouze s určitou „váhou“ – jistotou.

Nakonec přikládám klíč k indonéským výrazům (i když překlady nefungují v češtině stejně jako příslušná slova v indonéštině):

aduk	míchat	minum	pít
bumbui dengan	okořenit (s)	potong	krájet
didihkan	vařit	sisihkan	nechat odpočinout
makan	jíst	tambah	přidat
memeras	vymačkat	tuang	nalít
menghaluskan	rozmačkat		

air	voda	kelapa	kokos
anggur	víno	buah	ovoce
saus sambal	jakási omáčka	kecap manis	jakási jiná omáčka
adonan	těsto	cairan	kapalina
jeruk	citron	daging buah kelapa	kokosová dužina
kentang	brambora	bumbu	koření (zhruba)

Dovětek:

1. Po odhalení klíče si můžete domyslit, jak jsou data přikrášlená. Ne mnoho, to ne, ale aby byla úložka vůbec rozpletitelná, výskyty/nevýskyty jsem upravil, aby lépe seděly.

2. Rozlišovat slova podle toho, s jakými dalšími slovy se vyskytují, je v komputační sémantice široce praktikovaný způsob, jak zachytit význam slova. Tomuto přístupu se říká distributivní sémantika. Sice jsem výše napsal, že by naše metoda nefungovala moc dobře na větším množství dat, ale to jsme se snažili zjistit vztah pod-/nadřazenosti. Distributivní sémantika proti tomu funguje velmi dobře i na velkých datech, ve výsledku nám ovšem zpravidla jen řekne, která slova jsou si významově podobná, často aniž by rozlišovala jejich různé významy (resp. smysly).