Lexicography versus XML

Michal Měchura

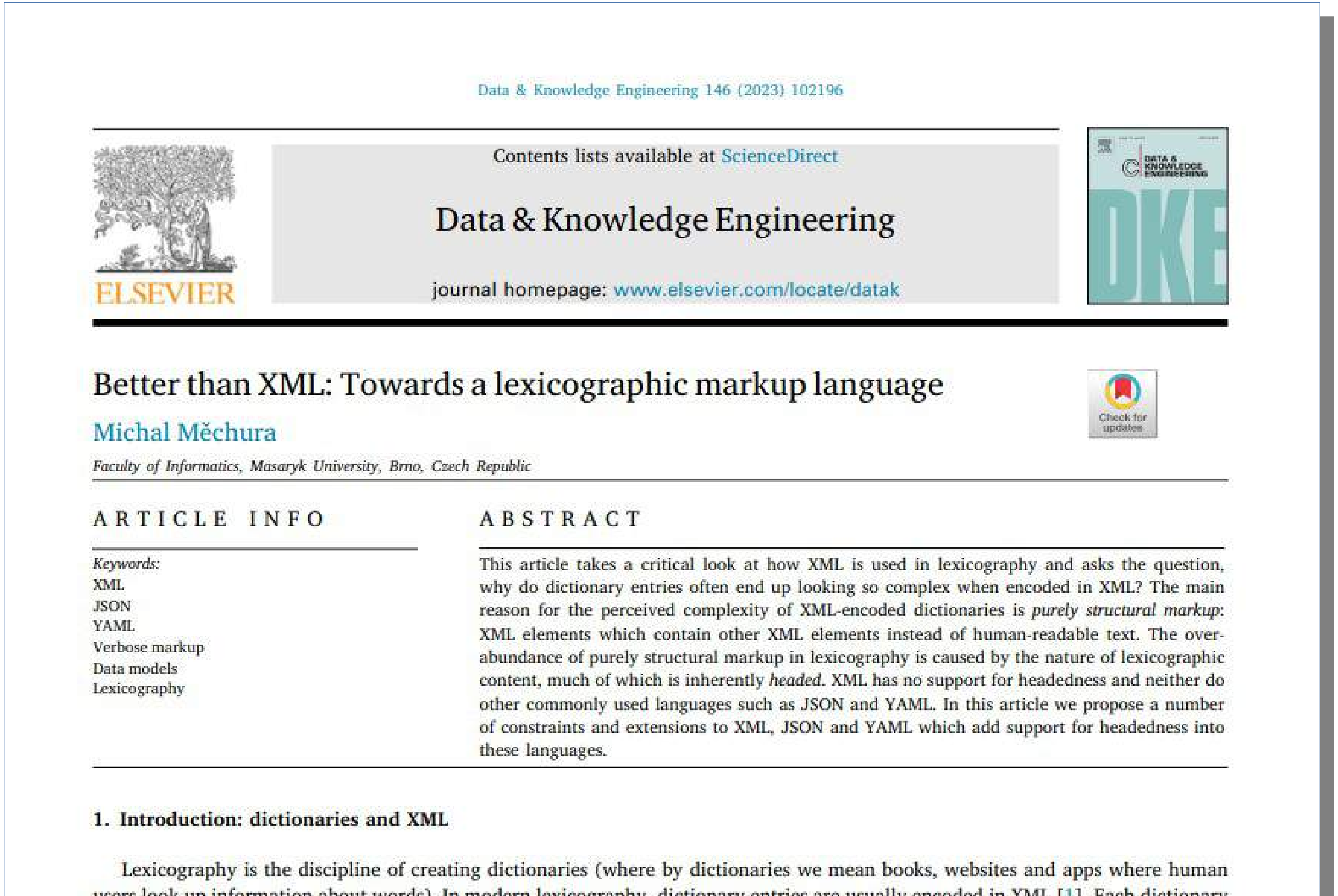
Masaryk University, Brno, Czech Republic

Declarative Amsterdam, November 2023

THE NEW
COLLINS GERMAN DICTIONARY
GERMAN
4TH edition
COLLINS SANSONI ITALIAN DICTIONARY
ITALIAN
THIRD EDITION
COLLINS ROBERT FRENCH DICTIONARY
FRENCH
Oxford Concise
Russian
New Internet Guide
New
Basic
Japanese-English
DICTIONARY
OXFORD
BONJINSHA
New
Pocket Oxford
Chinese
Dictionary
英汉·汉英
Talking Dictionary and Instant Translator
Only £19.99
COLLINS POCKET DICTIONARY
POLISH
new
Pocket Modern
Welsh
New
OXFORD
Oxford Starter Italian
New
OXFORD

ford
ncise
Russian
Basic
Japanese-English
DICTIONARY
OXFORD
BONJINSHA
New
Pocket
Oxford
Chinese
Dictionary
英汉·汉英
Talking
Dictionary
and
Instant
Translator
Only £19.99
New
COLLINS
POCKET
DICTIONARY
POLISH
new
Pocket
Modern
Welsh
New
OXFORD
Oxford
Starter
Italian
New
OXFORD

https://doi.org/10.1016/j.datak.2023.102196



Data & Knowledge Engineering 146 (2023) 102196

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Better than XML: Towards a lexicographic markup language

Michal Měchura

Faculty of Informatics, Masaryk University, Brno, Czech Republic

ARTICLE INFO

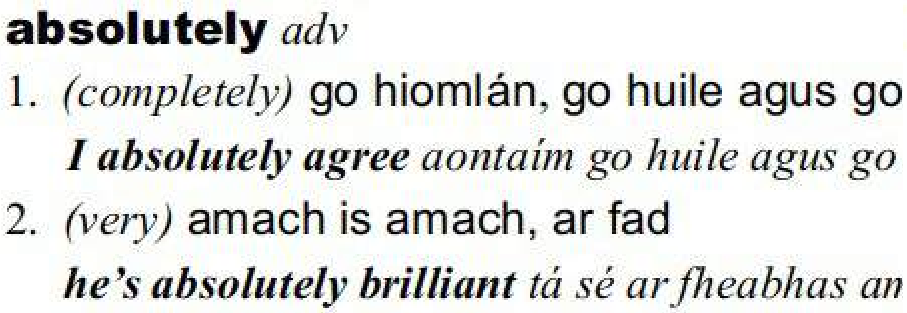
Keywords:
XML
JSON
YAML
Verbose markup
Data models
Lexicography

ABSTRACT

This article takes a critical look at how XML is used in lexicography and asks the question, why do dictionary entries often end up looking so complex when encoded in XML? The main reason for the perceived complexity of XML-encoded dictionaries is *purely structural markup*: XML elements which contain other XML elements instead of human-readable text. The over-abundance of purely structural markup in lexicography is caused by the nature of lexicographic content, much of which is inherently *headed*. XML has no support for headedness and neither do other commonly used languages such as JSON and YAML. In this article we propose a number of constraints and extensions to XML, JSON and YAML which add support for headedness into these languages.

1. Introduction: dictionaries and XML

Lexicography is the discipline of creating dictionaries (where by dictionaries we mean books, websites and apps where human users look up information about words). In modern lexicography, dictionary entries are usually encoded in XML [1]. Each dictionary



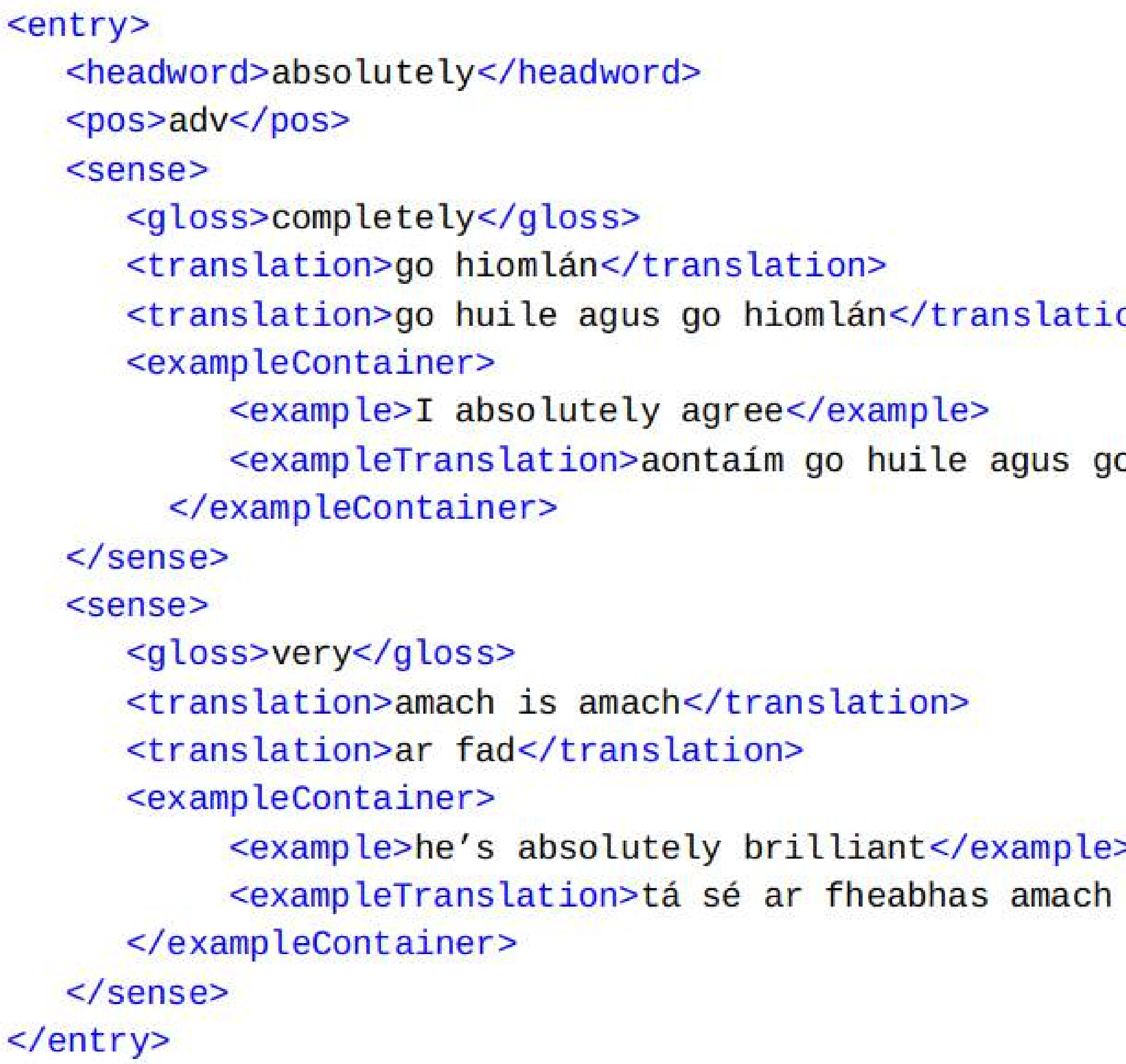
absolutely *adv*

1. *(completely)* go hiomlán, go huile agus go

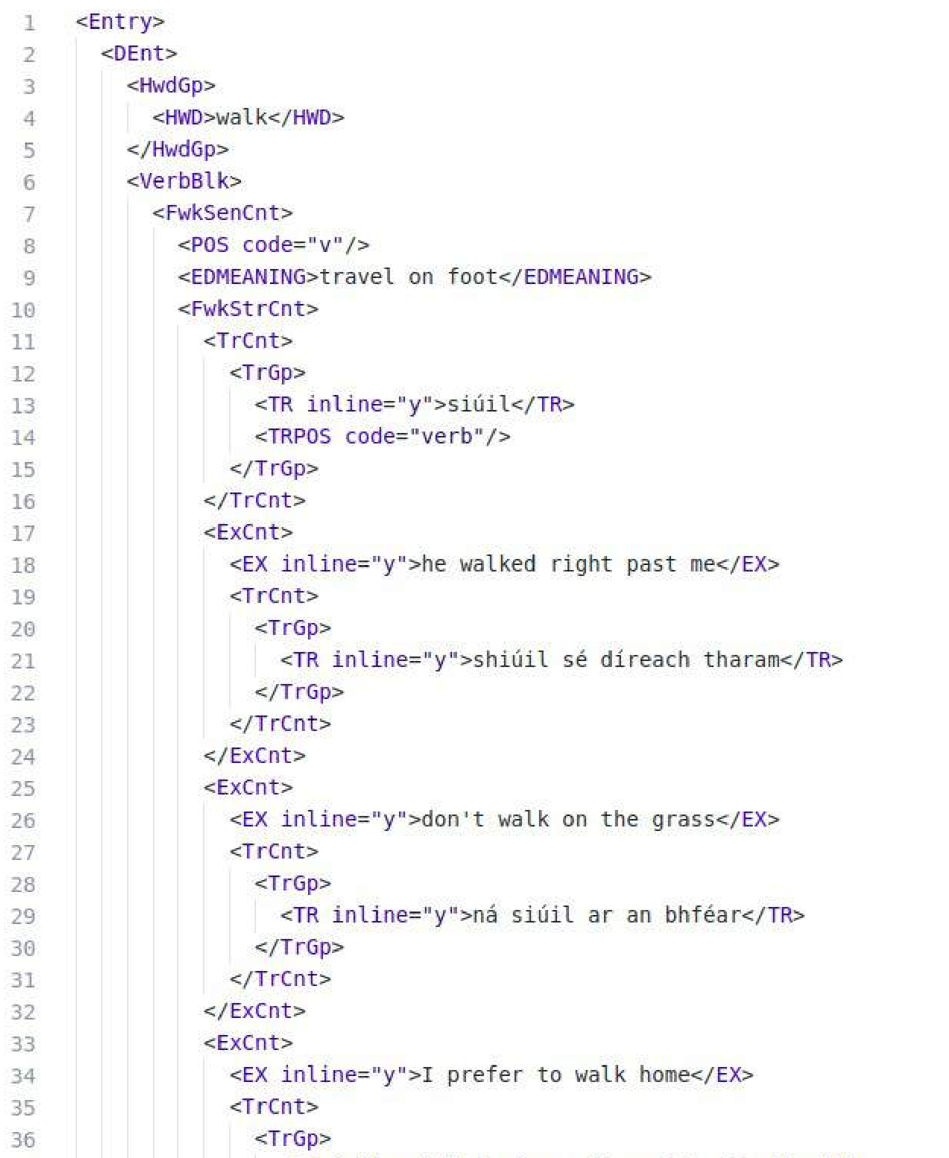
   I absolutely agree *aontaím go huile agus go*

2. *(very)* amach is amach, ar fad

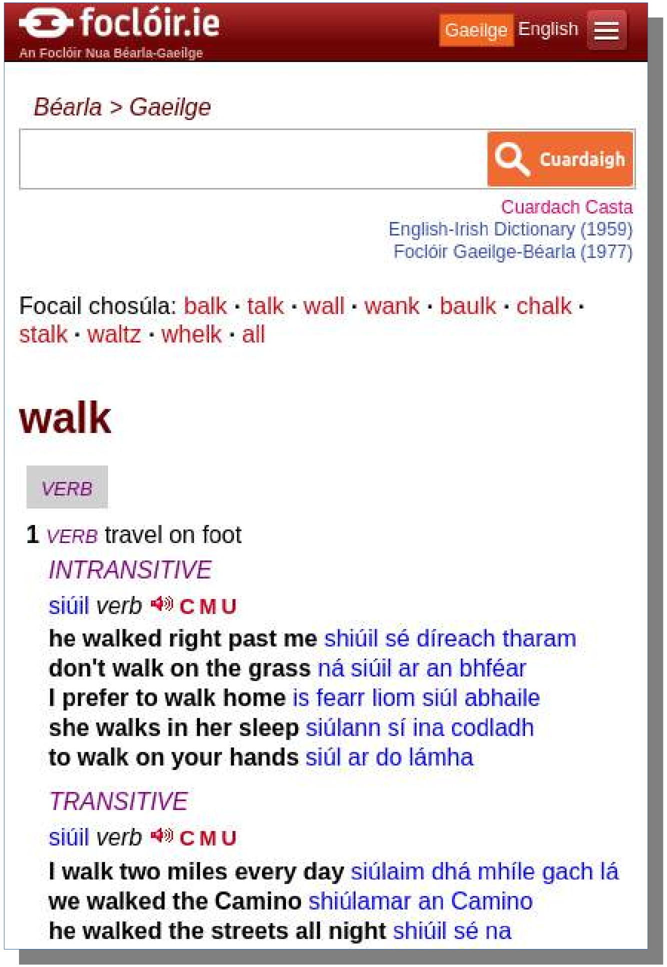
   he's absolutely brilliant *tá sé ar fheabhas an*



```
<entry>
   <headword>absolutely</headword>
   <pos>adv</pos>
   <sense>
      <gloss>completely</gloss>
      <translation>go hiomlán</translation>
      <translation>go huile agus go hiomlán</translatio
      <exampleContainer>
           <example>I absolutely agree</example>
           <exampleTranslation>aontaím go huile agus go
        </exampleContainer>
   </sense>
   <sense>
      <gloss>very</gloss>
      <translation>amach is amach</translation>
      <translation>ar fad</translation>
      <exampleContainer>
           <example>he’s absolutely brilliant</example>
           <exampleTranslation>tá sé ar fheabhas amach
      </exampleContainer>
   </sense>
</entry>
```



```
<Entry>
  <DEnt>
    <HwdGp>
      <HWD>walk</HWD>
    </HwdGp>
    <VerbBlk>
      <FwkSenCnt>
        <POS code="v"/>
        <EDMEANING>travel on foot</EDMEANING>
        <FwkStrCnt>
          <TrCnt>
            <TrGp>
              <TR inline="y">siúil</TR>
              <TRPOS code="verb"/>
            </TrGp>
          </TrCnt>
          <ExCnt>
            <EX inline="y">he walked right past me</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">shiúil sé díreach tharam</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">don't walk on the grass</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">ná siúil ar an bhféar</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">I prefer to walk home</EX>
            <TrCnt>
              <TrGp>
```



Béarla > Gaeilge

Cuardach Casta
English-Irish Dictionary (1959)
Foclóir Gaeilge-Béarla (1977)

Focail chosúla: balk · talk · wall · wank · baulk · chalk · stalk · waltz · whelk · all

walk

VERB

1 *VERB* travel on foot

INTRANSITIVE

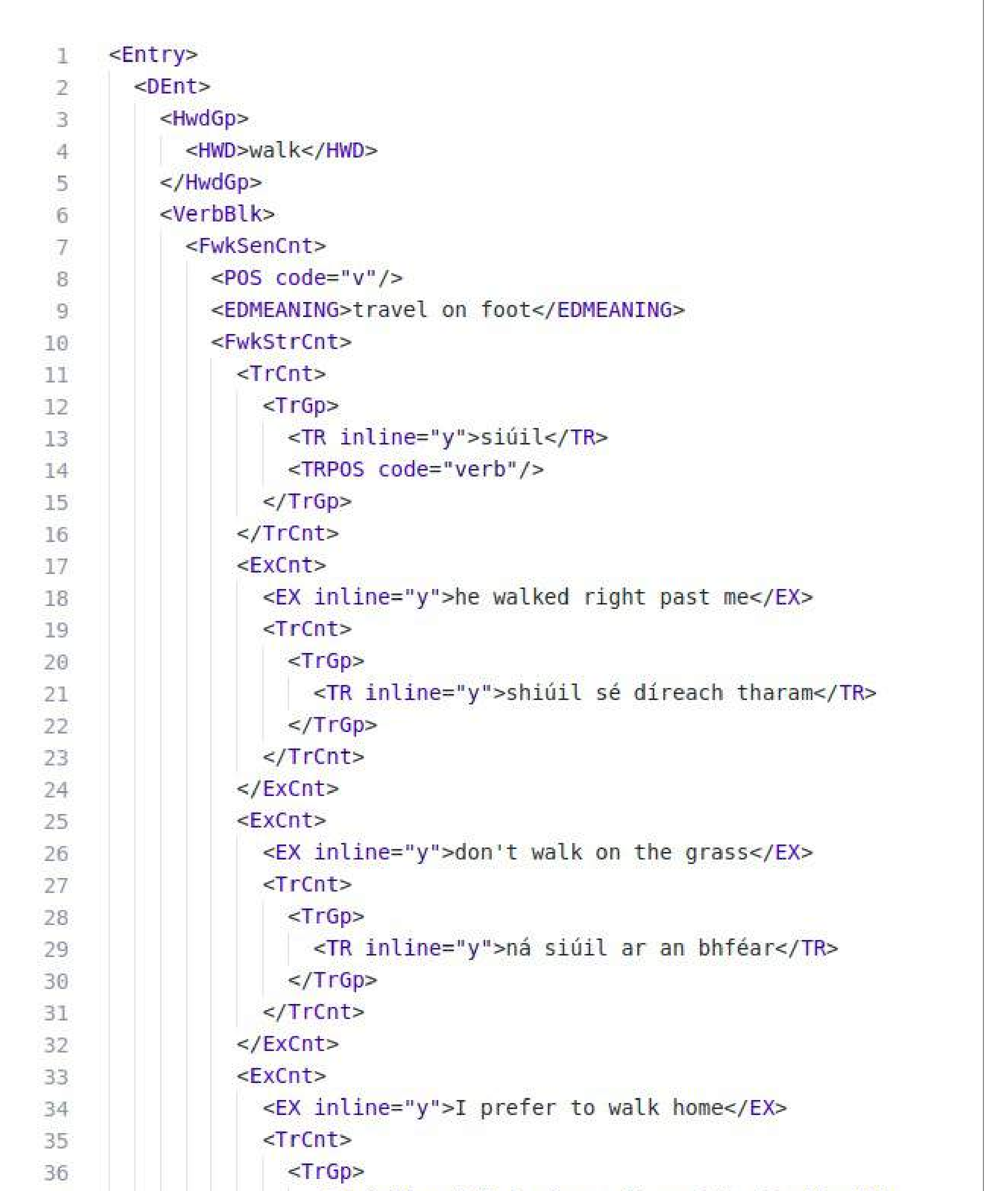
siúil *verb* C M U

he walked right past me shiúil sé díreach tharam
don't walk on the grass ná siúil ar an bhféar
I prefer to walk home is fearr liom siúl abhaile
she walks in her sleep siúlann sí ina codladh
to walk on your hands siúl ar do lámha

TRANSITIVE

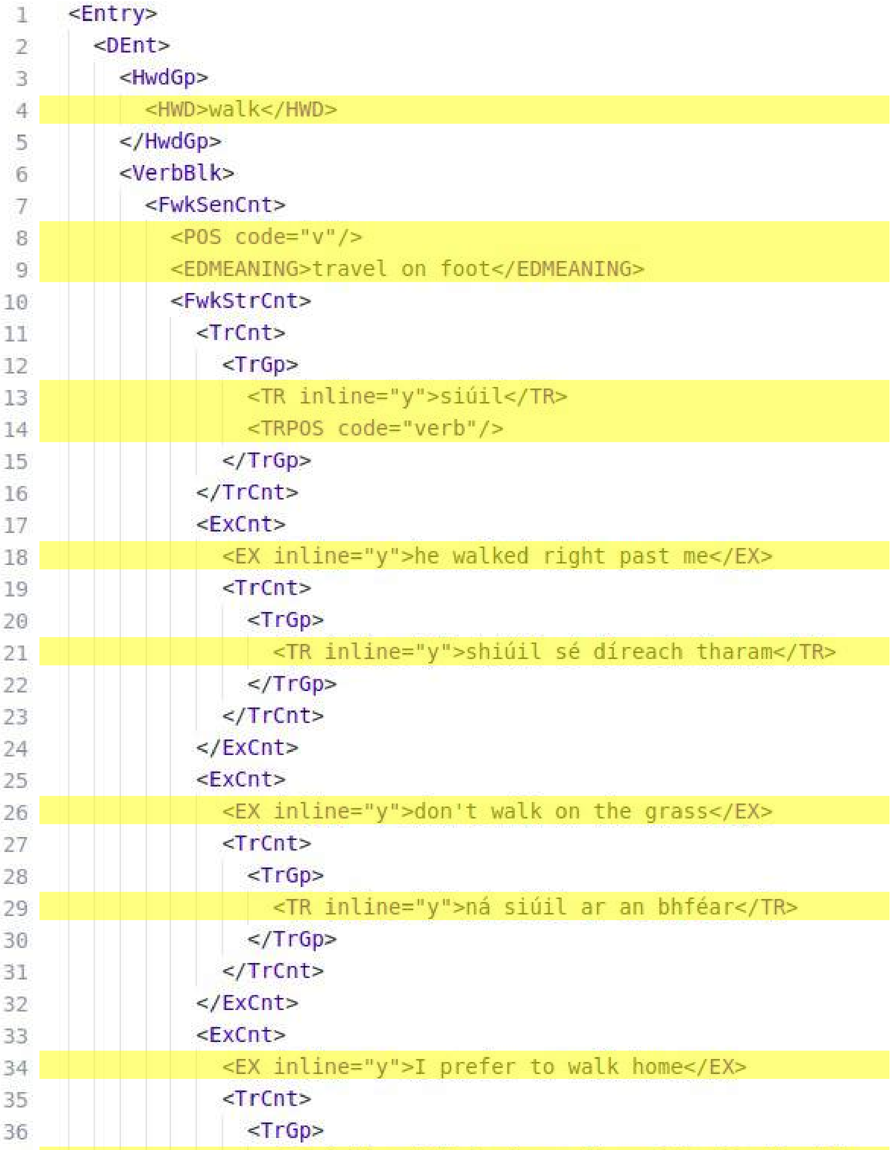
siúil *verb* C M U

I walk two miles every day siúlaim dhá mhíle gach lá
we walked the Camino shiúlamar an Camino
he walked the streets all night shiúil sé na



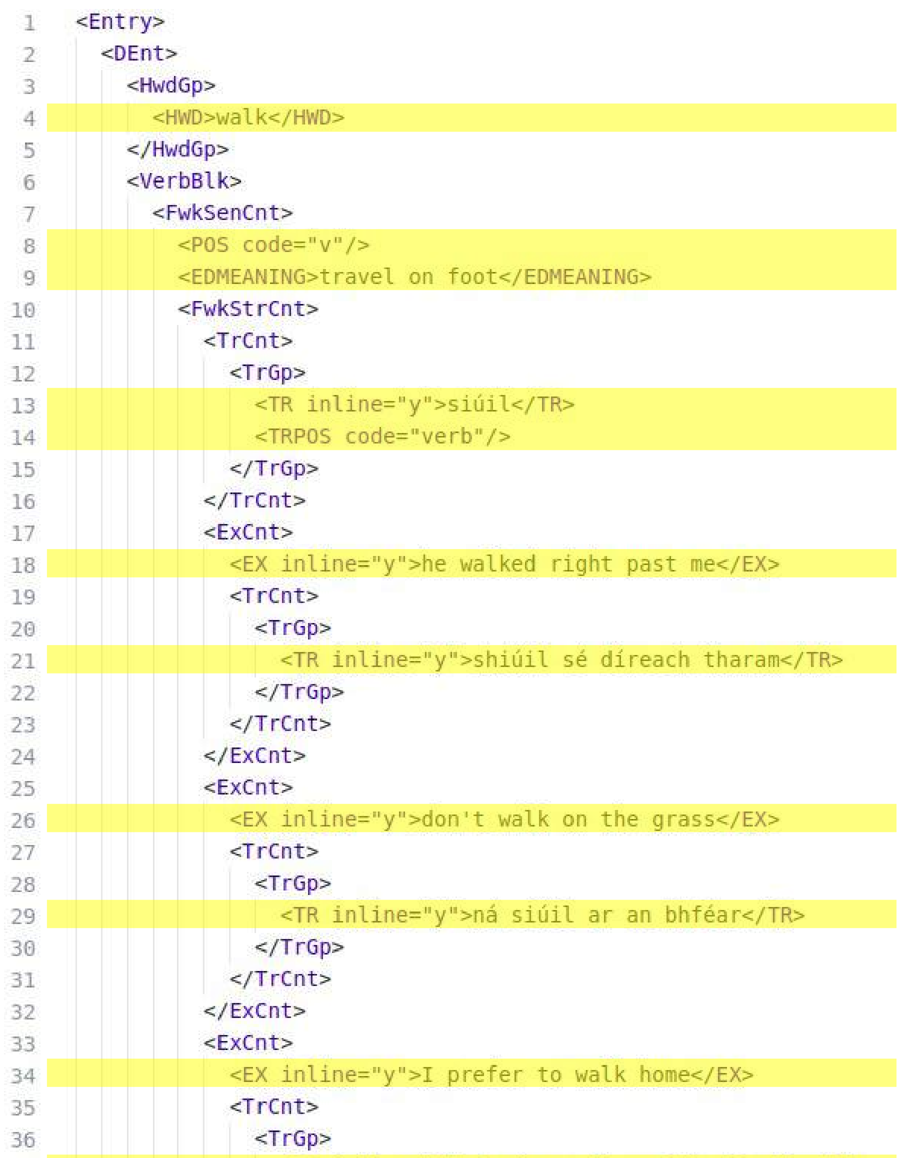
```
<Entry>
  <DEnt>
    <HwdGp>
      <HWD>walk</HWD>
    </HwdGp>
    <VerbBlk>
      <FwkSenCnt>
        <POS code="v"/>
        <EDMEANING>travel on foot</EDMEANING>
        <FwkStrCnt>
          <TrCnt>
            <TrGp>
              <TR inline="y">siúil</TR>
              <TRPOS code="verb"/>
            </TrGp>
          </TrCnt>
          <ExCnt>
            <EX inline="y">he walked right past me</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">shiúil sé díreach tharam</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">don't walk on the grass</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">ná siúil ar an bhféar</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">I prefer to walk home</EX>
            <TrCnt>
              <TrGp>
```

“Matryoshkization”



```
<Entry>
  <DEnt>
    <HwdGp>
      <HWD>walk</HWD>
    </HwdGp>
    <VerbBlk>
      <FwkSenCnt>
        <POS code="v"/>
        <EDMEANING>travel on foot</EDMEANING>
        <FwkStrCnt>
          <TrCnt>
            <TrGp>
              <TR inline="y">siúil</TR>
              <TRPOS code="verb"/>
            </TrGp>
          </TrCnt>
          <ExCnt>
            <EX inline="y">he walked right past me</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">shiúil sé díreach tharam</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">don't walk on the grass</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">ná siúil ar an bhféar</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">I prefer to walk home</EX>
            <TrCnt>
              <TrGp>
```

“Matryoshkization”

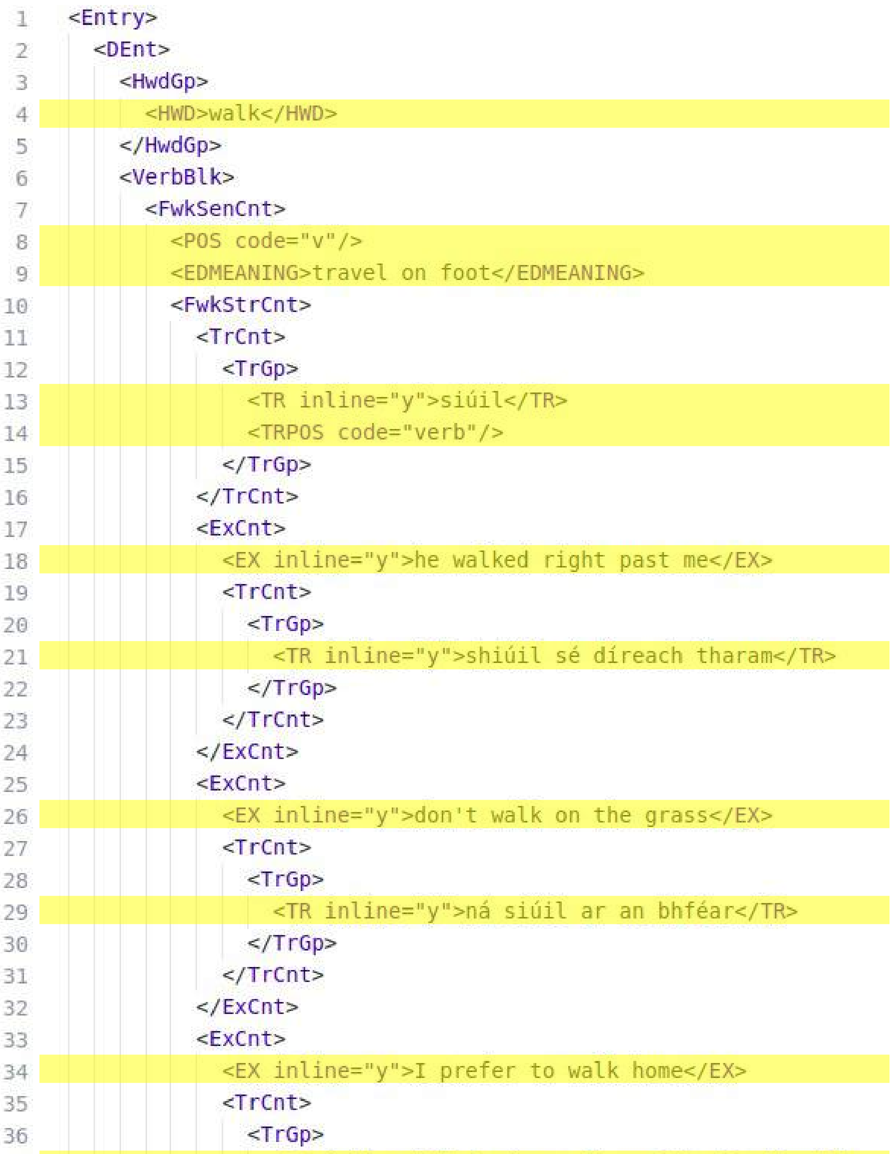


```
<Entry>
  <DEnt>
    <HwdGp>
      <HWD>walk</HWD>
    </HwdGp>
    <VerbBlk>
      <FwkSenCnt>
        <POS code="v"/>
        <EDMEANING>travel on foot</EDMEANING>
        <FwkStrCnt>
          <TrCnt>
            <TrGp>
              <TR inline="y">siúil</TR>
              <TRPOS code="verb"/>
            </TrGp>
          </TrCnt>
          <ExCnt>
            <EX inline="y">he walked right past me</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">shiúil sé díreach tharam</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">don't walk on the grass</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">ná siúil ar an bhféar</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">I prefer to walk home</EX>
            <TrCnt>
              <TrGp>
```

2,387 lines of code

- **957** (40%) human-readable text
- **1,430** (60%) purely structural markup

“Matryoshkization”



```
<Entry>
  <DEnt>
    <HwdGp>
      <HWD>walk</HWD>
    </HwdGp>
    <VerbBlk>
      <FwkSenCnt>
        <POS code="v"/>
        <EDMEANING>travel on foot</EDMEANING>
        <FwkStrCnt>
          <TrCnt>
            <TrGp>
              <TR inline="y">siúil</TR>
              <TRPOS code="verb"/>
            </TrGp>
          </TrCnt>
          <ExCnt>
            <EX inline="y">he walked right past me</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">shiúil sé díreach tharam</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">don't walk on the grass</EX>
            <TrCnt>
              <TrGp>
                <TR inline="y">ná siúil ar an bhféar</TR>
              </TrGp>
            </TrCnt>
          </ExCnt>
          <ExCnt>
            <EX inline="y">I prefer to walk home</EX>
            <TrCnt>
              <TrGp>
```

2,387 lines of code

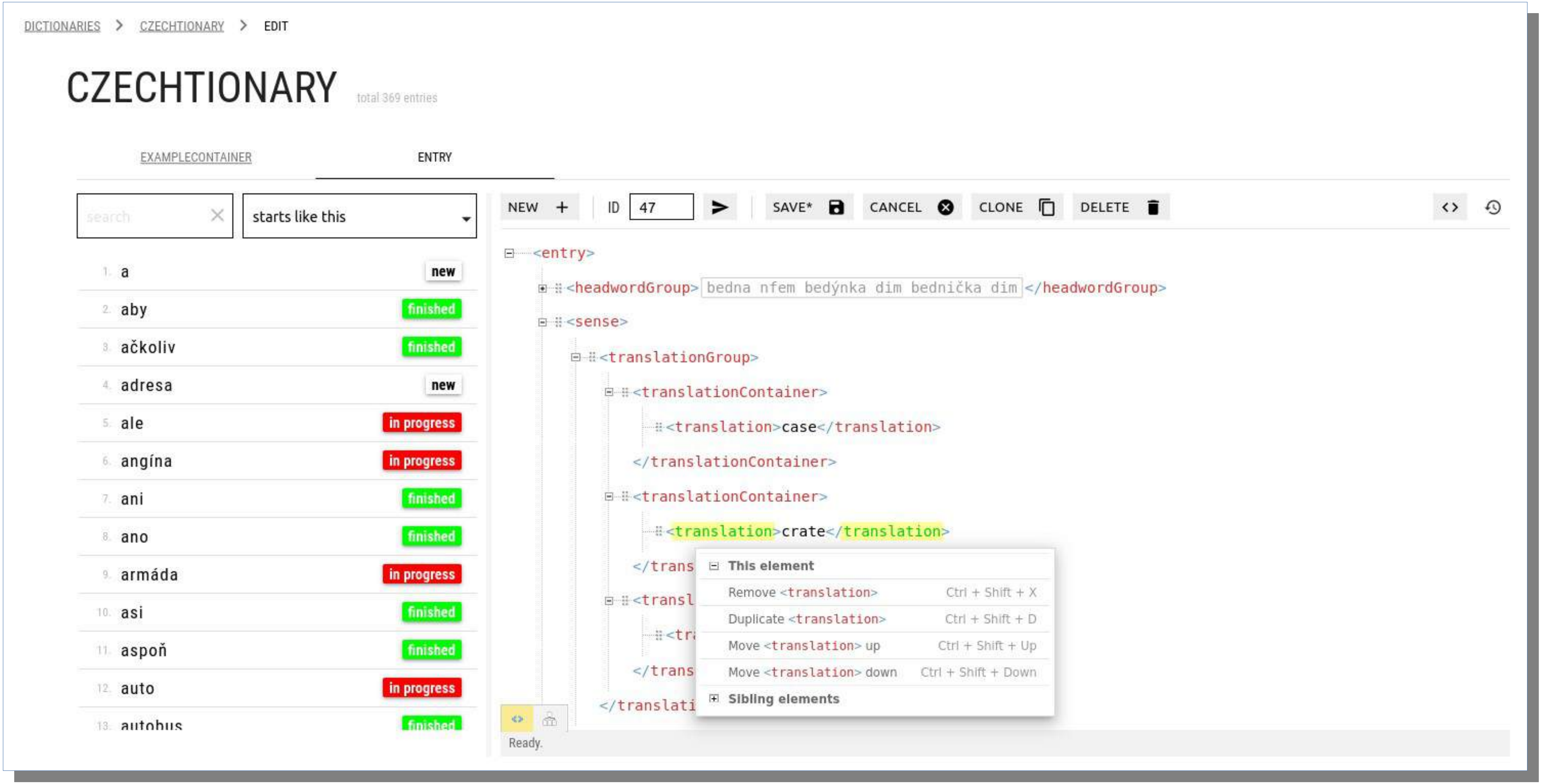
- **957** (40%) human-readable text
- **1,430** (60%) purely structural markup

1,672 elements

- **957** (57%) human-readable text
- **715** (43%) purely structural markup

“Matryoshkization”

Is matryoshkization really such a big problem?


DICTIONARIES > CZECHTIONARY > EDIT
CZECHTIONARY
total 369 entries
EXAMPLECONTAINER
ENTRY
search
starts like this
NEW +
ID 47
SAVE*
CANCEL
CLONE
DELETE
This element
Remove <translation> Ctrl + Shift + X
Duplicate <translation> Ctrl + Shift + D
Move <translation> up Ctrl + Shift + Up
Move <translation> down Ctrl + Shift + Down
Sibling elements
Ready.

1.	a	new
2.	aby	finished
3.	ačkoliv	finished
4.	adresa	new
5.	ale	in progress
6.	angína	in progress
7.	ani	finished
8.	ano	finished
9.	armáda	in progress
10.	asi	finished
11.	aspoň	finished
12.	auto	in progress
13.	autobus	finished

```
<entry>
  <headwordGroup>bedna nfem bedýnka dim bednička dim</headwordGroup>
  <sense>
    <translationGroup>
      <translationContainer>
        <translation>case</translation>
      </translationContainer>
      <translationContainer>
        <translation>crate</translation>
      </trans
      <transl
        <tr
      </trans
    </translati
```

> “Notations affect what you can do with them.”
>
> — Steven Pemberton, ‘On the Descriptions of Data:
> The Usability of Notations’, XML Prague 2017

$$\begin{array}{r} \text{XVII} \\ +\ \ \text{IV} \\ \hline \text{XXI} \end{array}$$

$$\begin{array}{r} 17 \\ +\ \ 4 \\ \hline 21 \end{array}$$

Schema migration

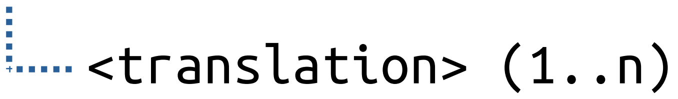
```
<translation> (1..n)
```

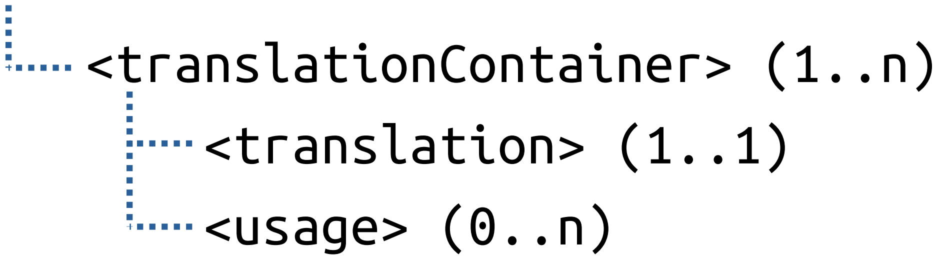
Schema migration

```
<translation> (1..n)
```

```
<translation>leasú</translation>
<translation>athchóiriú</translation>
```

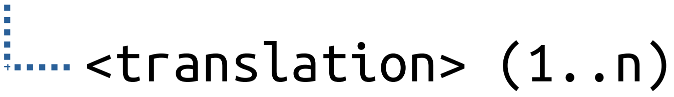
Schema migration



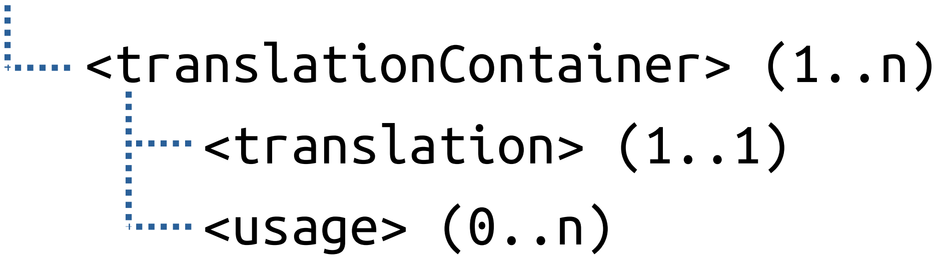


```
<translation>leasú</translation>
<translation>athchóiriú</translation>
```

Schema migration



<translation> (1..n)



<translationContainer> (1..n)

<translation> (1..1)

<usage> (0..n)

```
<translation>leasú</translation>
<translation>athchóiriú</translation>
```

```
<translationContainer>
  <translation>leasú</translation>
</translationContainer>
<translationContainer>
  <translation>athchóiriú</translation>
</translationContainer>
```

Can matryoshkization be avoided?

The ‘head + modifiers’ design pattern

```
<translationGroup>
  <translation>athchóiriú</translation>
  <pos>n-masc</pos>
  <usage>formal</usage>
</translationGroup>
```

The ‘head + modifiers’ design pattern

```
<translationGroup>
  <translation>athchóiriú</translation>
  <pos>n-masc</pos>
  <usage>formal</usage>
</translationGroup>
```

The ‘head + modifiers’ design pattern

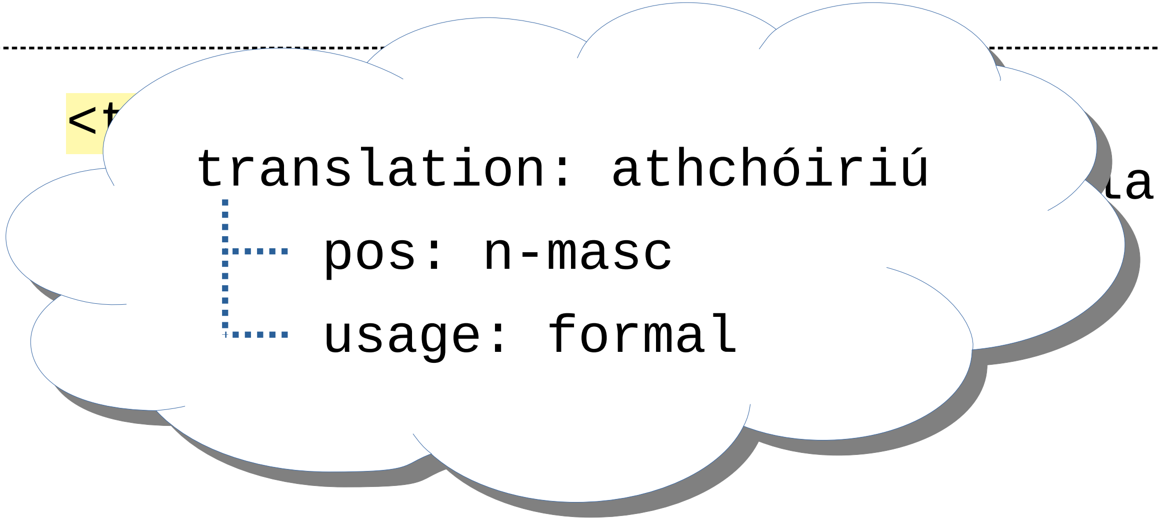
```
<translationGroup>
  <translation>athchóiriú</translation>
  <pos>n-masc</pos>
  <usage>formal</usage>
</translationGroup>
```

The ‘head + modifiers’ design pattern

```
<translationGroup>
  <translation>athchóiriú</translation>
  <pos>n-masc</pos>
  <usage>formal</usage>
</translationGroup>
```

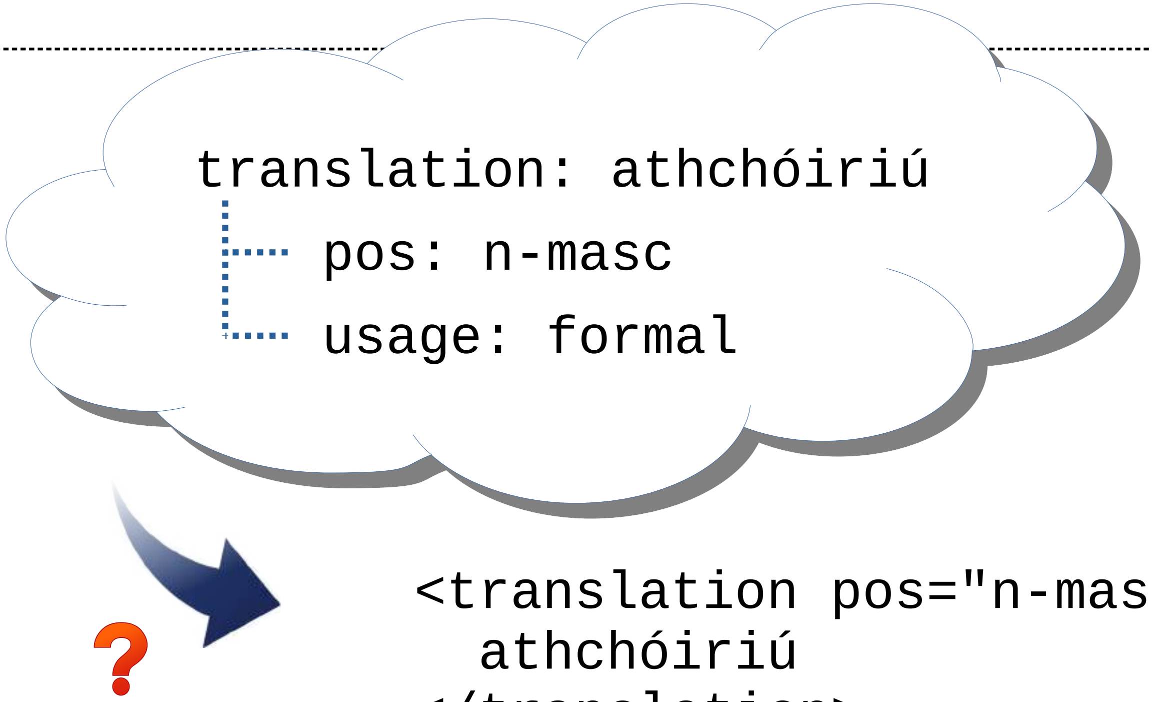
The ‘head + modifiers’ design pattern

```
<t                                      lation>
```



```
translation: athchóiriú
  ├── pos: n-masc
  └── usage: formal
```

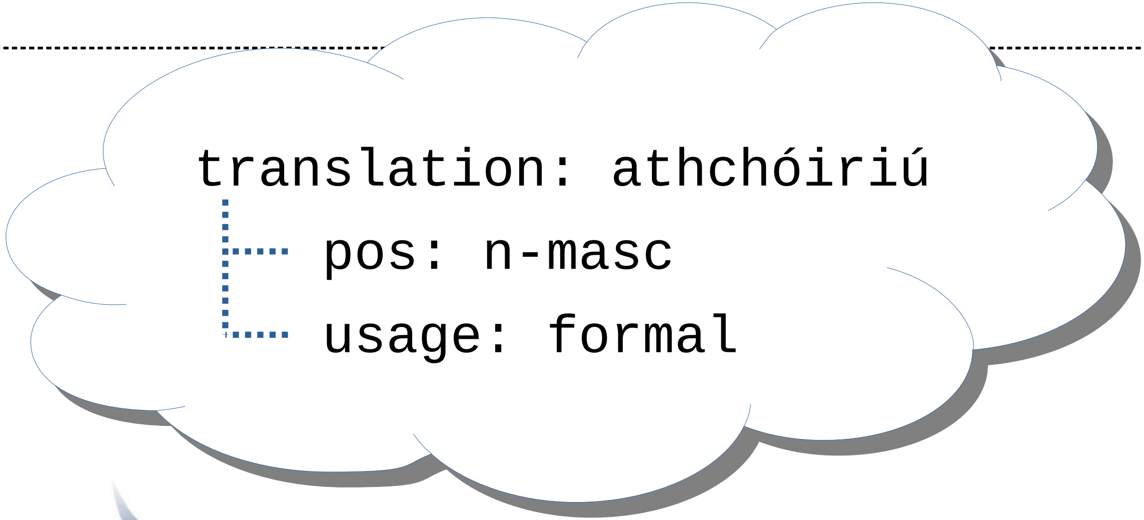
Strategy 1: modifiers as attributes



```
translation: athchóiriú
 ├── pos: n-masc
 └── usage: formal
```

```
<translation pos="n-masc" usage="formal">
  athchóiriú
</translation>
```

Strategy 1: modifiers as attributes

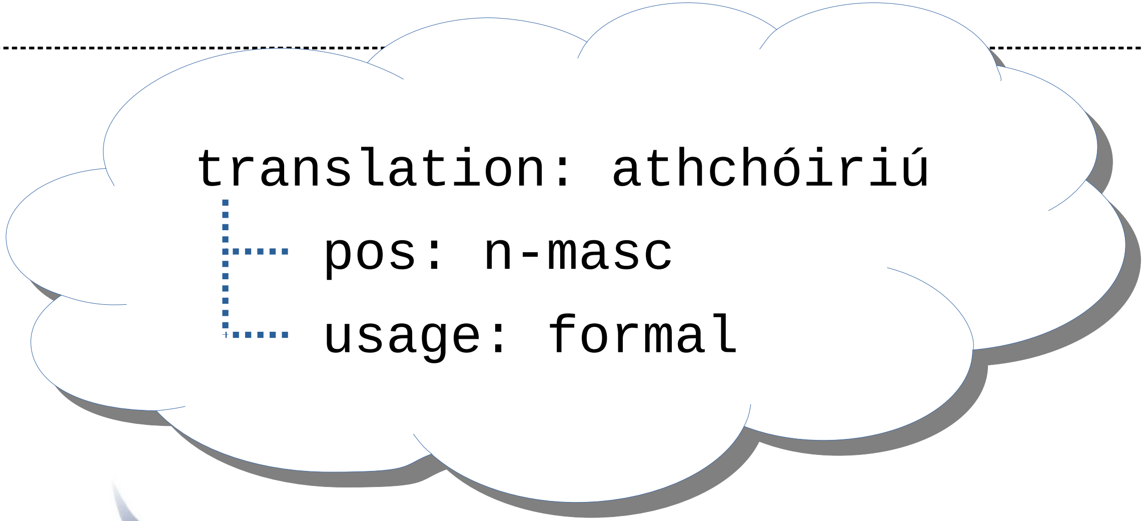


```
translation: athchóiriú
   pos: n-masc
   usage: formal
```

?

```
<translation pos="n-masc" usage="formal">
  athchóiriú
</translation>
```

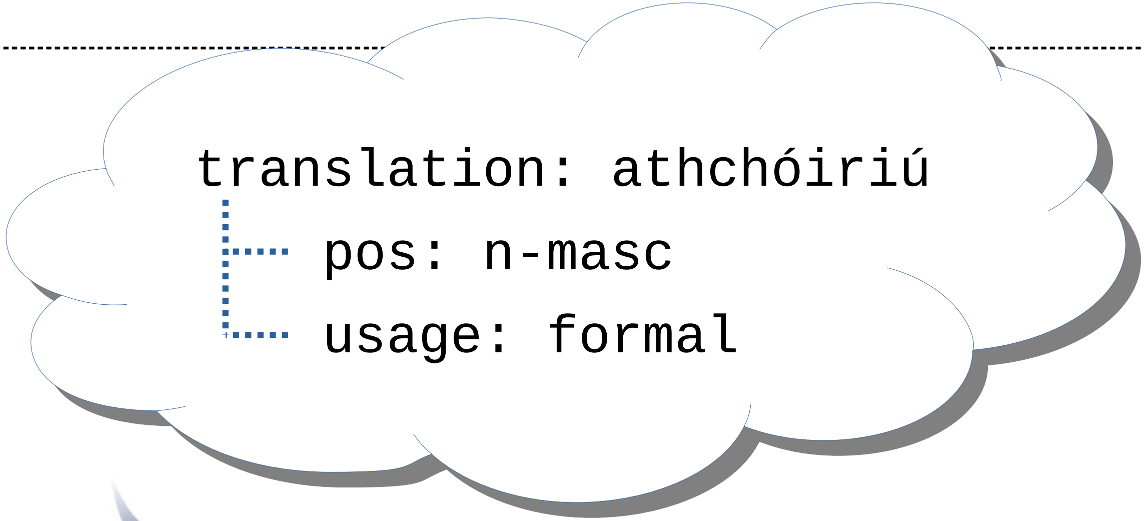
Strategy 1: modifiers as attributes



```
translation: athchóiriú
   ├── pos: n-masc
   └── usage: formal
```

```
<translation pos="n-masc" usage="formal">
  athchóiriú
</translation>
```

Strategy 2: head as attribute



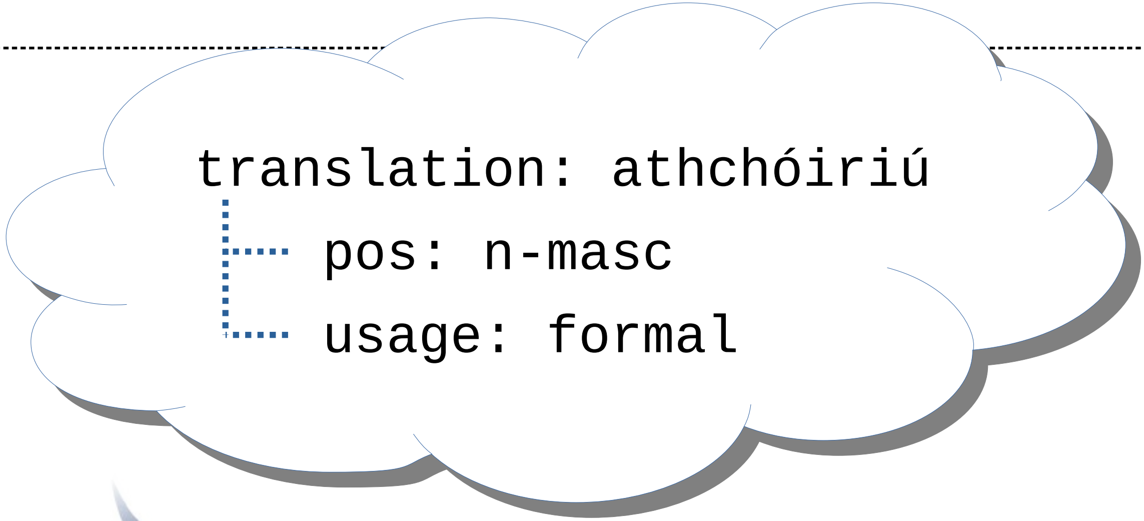
translation: athchóiriú

pos: n-masc

usage: formal

```
<translation value="athchóiriú">
  <pos>n-masc</pos>
  <usage>formal</usage>
</translation>
```

Strategy 2: head as attribute



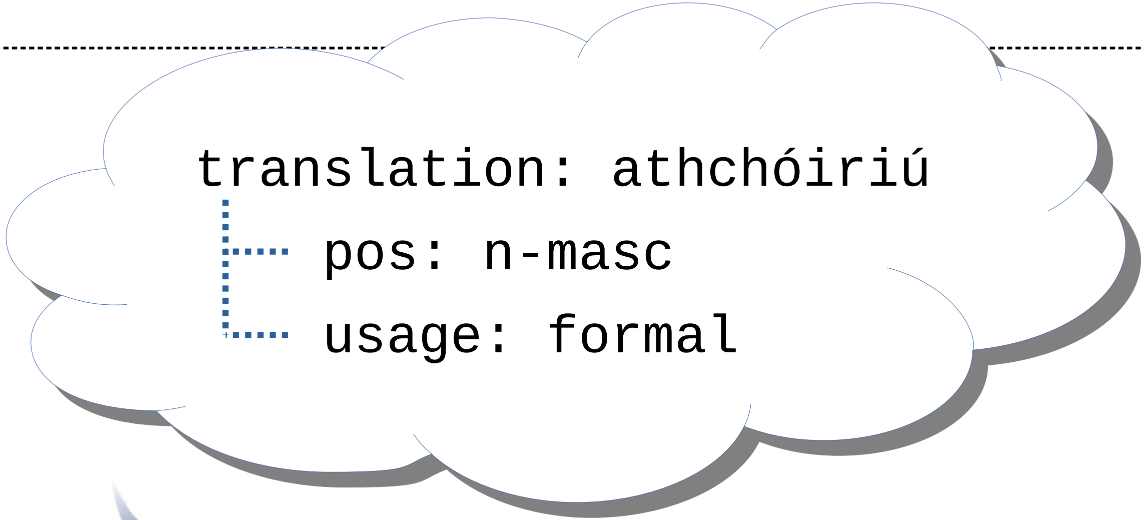
translation: athchóiriú

pos: n-masc

usage: formal

```
<translation value="athchóiriú">
  <pos>n-masc</pos>
  <usage>formal</usage>
</translation>
```

Strategy 2: head as attribute



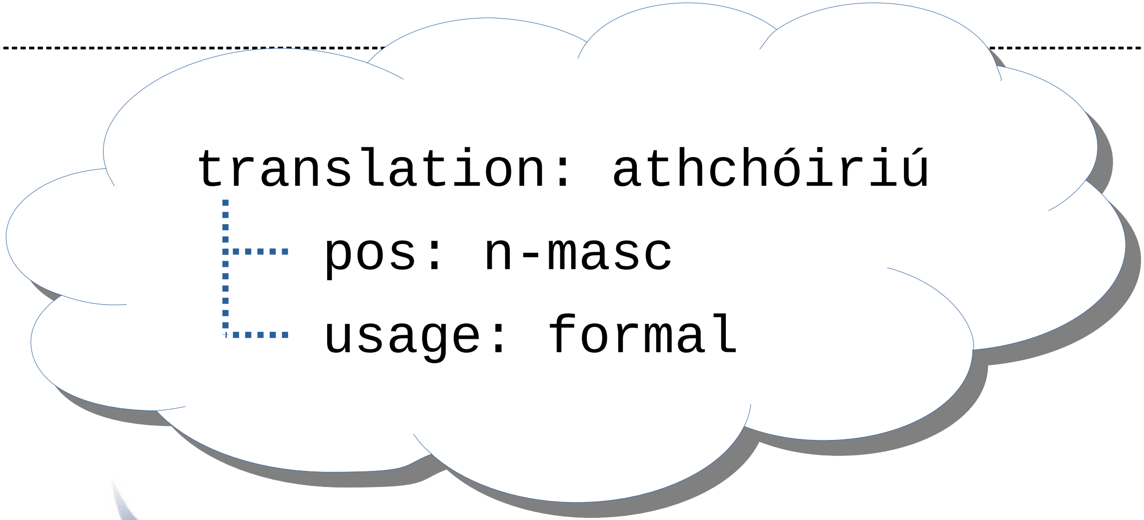
translation: athchóiriú

pos: n-masc

usage: formal

```
<translation value="athchóiriú">
  <pos>n-masc</pos>
  <usage>formal</usage>
</translation>
```

Strategy 3: mixed content



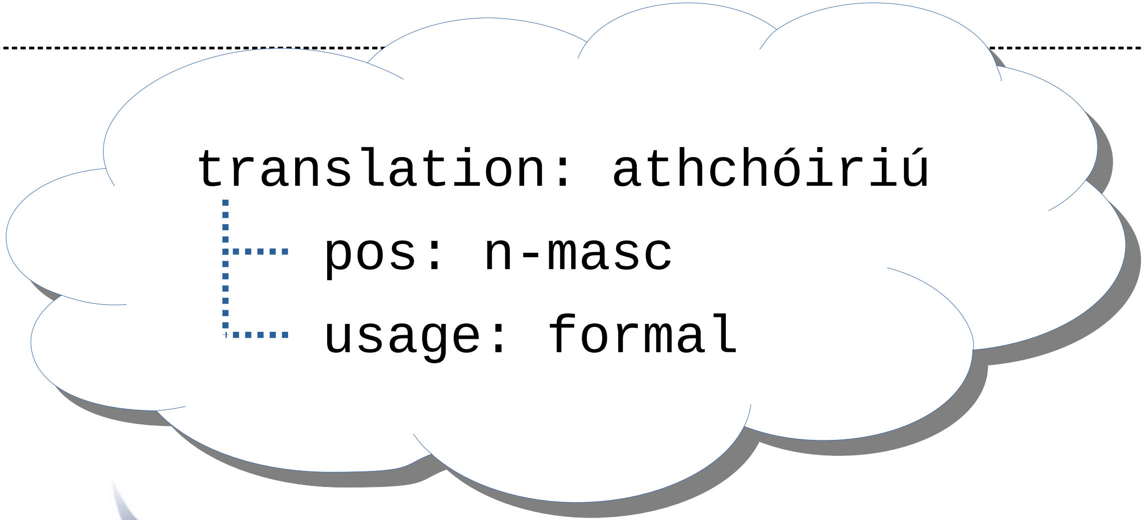
translation: athchóiriú

pos: n-masc

usage: formal

```
<translation>
  athchóiriú
  <pos>n-masc</pos>
  <usage>formal</usage>
</translation>
```

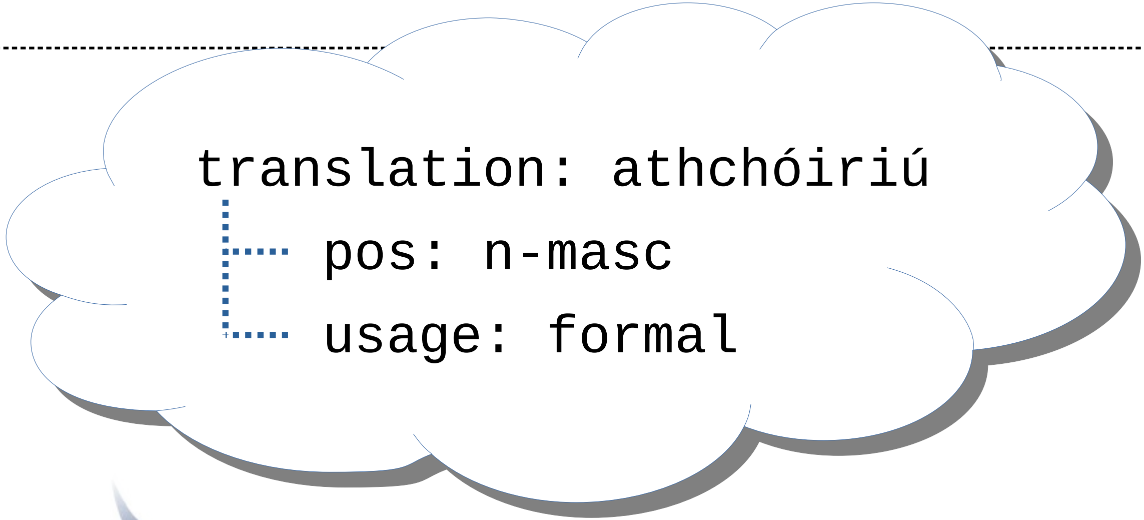
Strategy 3: mixed content



```
translation: athchóiriú
    pos: n-masc
    usage: formal
```

```
<translation>
  athchóiriú
  <pos>n-masc</pos>
  <usage>formal</usage>
</translation>
```

Strategy 3: mixed content

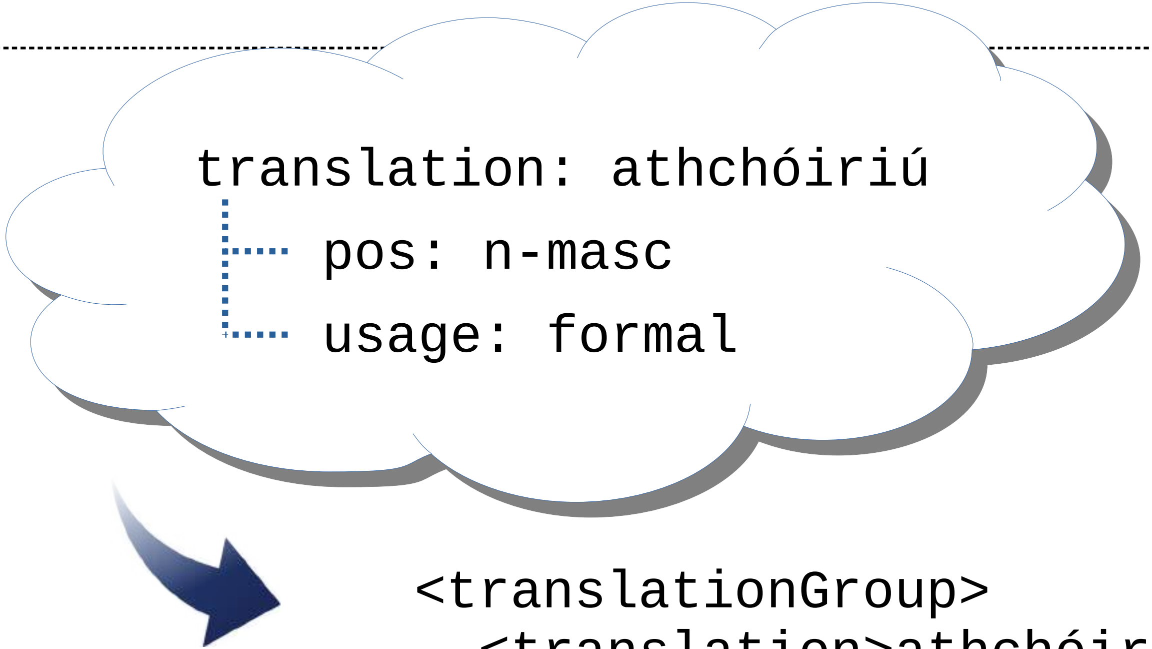


translation: athchóiriú

- pos: n-masc
- usage: formal

```
<translation>
  athchóiriú
  <pos>n-masc</pos>
  <usage>formal</usage>
</translation>
```

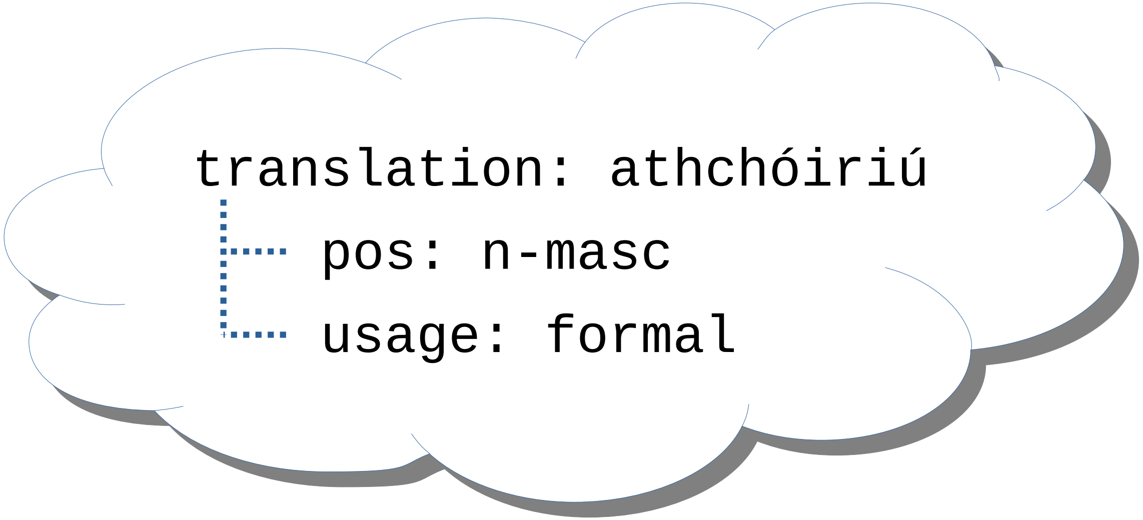
Strategy 4: matryoshkization



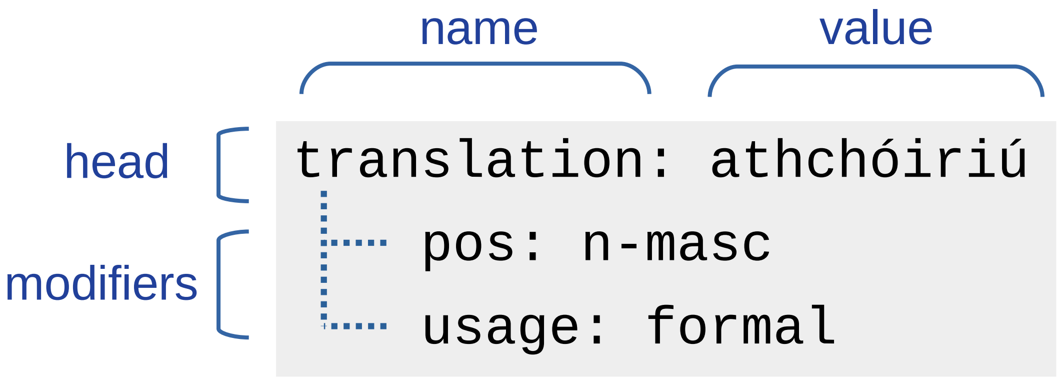
```
translation: athchóiriú
  pos: n-masc
  usage: formal
```

```
<translationGroup>
  <translation>athchóiriú</translation>
  <pos>n-masc</pos>
  <usage>formal</usage>
</translationGroup>
```

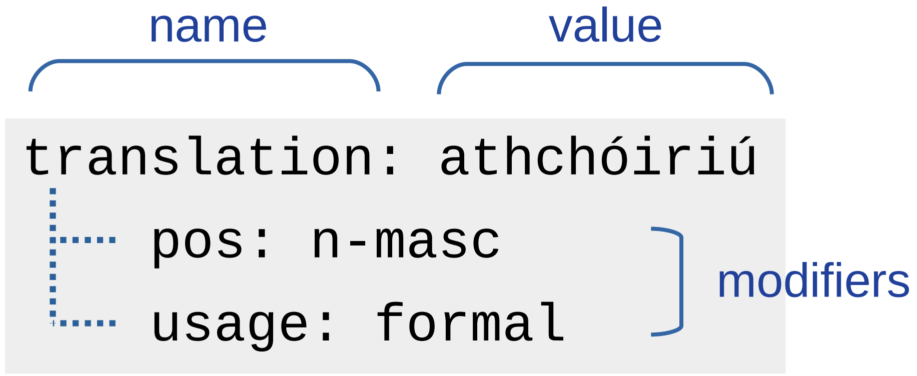
Why are headed structures so difficult to represent in XML?



```
translation: athchóiriú
  ├── pos: n-masc
  └── usage: formal
```

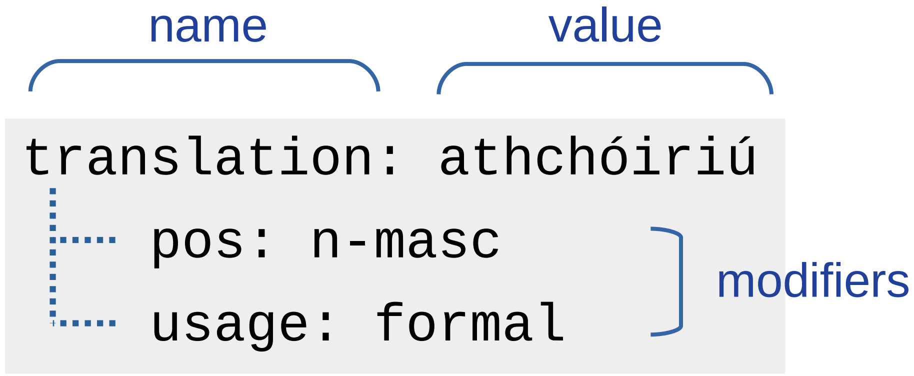


```
translation: athchóiriú
    pos: n-masc
    usage: formal
```



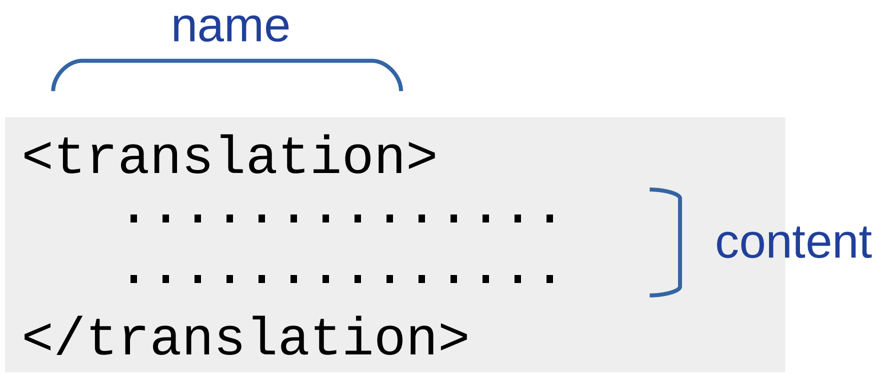
```
translation: athchóiriú
    pos: n-masc
    usage: formal
```

<name, value, modifiers>



```
translation: athchóiriú
    pos: n-masc
    usage: formal
```

<name, value, modifiers>



```
<translation>
   ..............
   ..............
</translation>
```

<name, content>

Looking outside XML

JSON

```
{
    "headword": "bear",
    "pos": "noun",
    "senses": [{
        "definition": "an animal which...",
        "example": "watch out there are bears..."
    }, {
        "definition": "a person who..."
    }]
}
```

JSON

```
{
    "headword": "bear",
    "pos": "noun",
    "senses": [{
        "definition": "an animal which...",
        "example": "watch out there are bears..."
    }, {
        "definition": "a person who..."
    }]
}
```

JSON

```
{
    "headword": "bear",
    "pos": "noun",
    "senses": [{
        "definition": "an animal which...",
        "example": "watch out there are bears..."
    }, {
        "definition": "a person who..."
    }]
}
```

JSON

```
{
    "headword": "bear",
    "pos": "noun",
    "senses": [{
        "definition": "an animal which...",
        "example": "watch out there are bears..."
    }, {
        "definition": "a person who..."
    }]
}
```

JSON

```
{
    "headword": "bear",
    "pos": "noun",
    "senses": [{
        "definition": "an animal which...",
        "example": "watch out there are bears..."
    }, {
        "definition": "a person who..."
    }]
}
```

JSON

```
{
    "headword": "bear",
    "pos": "noun",
    "senses": [{
        "definition": "an animal which...",
        "example": "watch out there are bears..."
    }, {
        "definition": "a person who..."
    }]
}
```

YAML

```yaml
entry:
    headword: bear
    pos: noun
    senses:
        - definition: an animal which...
          example: watch out there are bears...
        - definition: a person who...
```

YAML

```
entry:
    headword: bear
    pos: noun
    senses:
        - definition: an animal which...
          example: watch out there are bears...
        - definition: a person who...
```

YAML

```
entry:
    headword: bear
    pos: noun
    senses:
        - definition: an animal which...
          example: watch out there are bears...
        - definition: a person who...
```

YAML

```
entry:
    headword: bear
    pos: noun
    senses:
        - definition: an animal which...
          example: watch out there are bears...
        - definition: a person who...
```

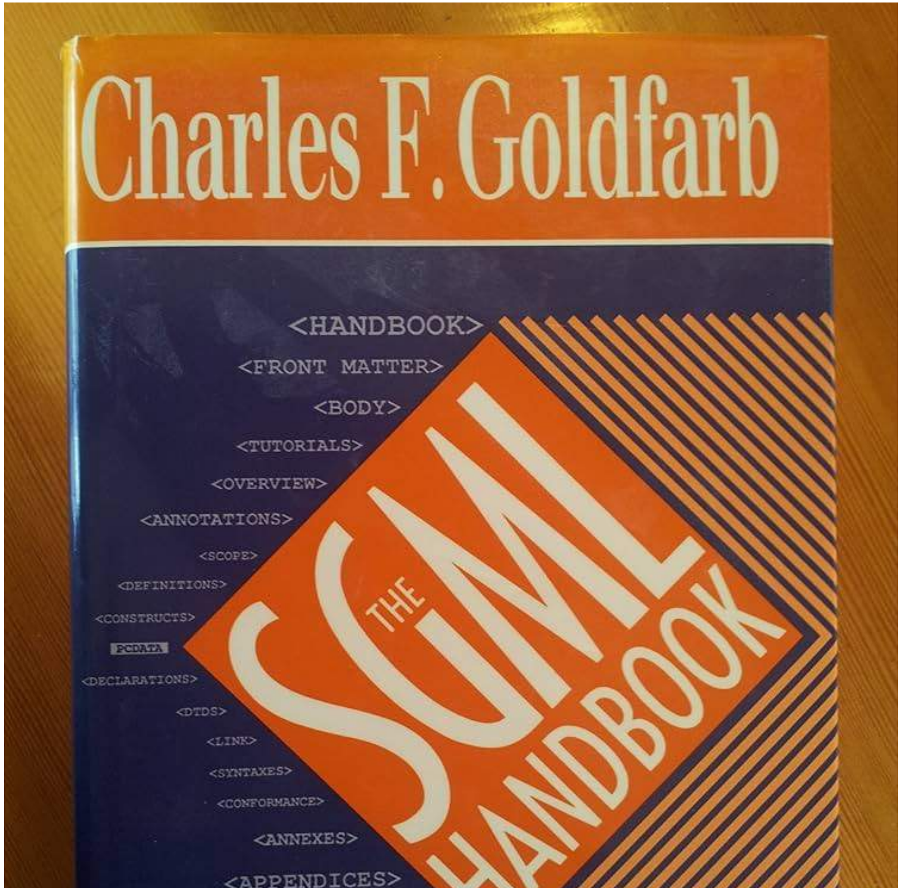
YAML

```
entry:
    headword: bear
    pos: noun
    senses:
        - definition: an animal which...
          example: watch out there are bears...
        - definition: a person who...
```

YAML

```
entry:
    headword: bear
    pos: noun
    senses:
        - definition: an animal which...
          example: watch out there are bears...
        - definition: a person who...
```

SGML



SGML: markup minimization

```
...
  <translation>athchóiriú
  <pos>n-masc
...
```

SGML: markup minimization

```
...
  <translation>athchóiriú</translation>
  <pos>n-masc</pos>
...
```

SGML: implicit elements

```
<translation>
  <value>athchóiriú</value>
  <pos>n-masc</pos>
  <usage>formal</usage>
<translation>
```

SGML: implicit elements

```
<translation>
  <value>athchóiriú</value>
  <pos>n-masc</pos>
  <usage>formal</usage>
<translation>
```

SGML: implicit elements

```
<translation>
  athchóiriú
  <pos>n-masc</pos>
  <usage>formal</usage>
<translation>
```

SGML: schema migration

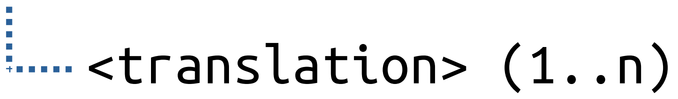
```
<translation> (1..n)
```

SGML: schema migration

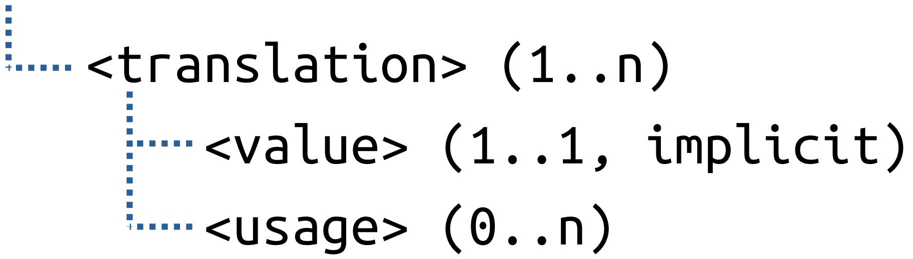
```
<translation> (1..n)
```

```
<translation>leasú</translation>
<translation>athchóiriú</translation>
```

SGML: schema migration



<translation> (1..n)



<translation> (1..n)

<value> (1..1, implicit)

<usage> (0..n)

```
<translation>leasú</translation>
<translation>athchóiriú</translation>
```

SGML: schema migration

```
<translation> (1..n)
```

```
<translation> (1..n)
    <value> (1..1, implicit)
    <usage> (0..n)
```

```
<translation>leasú</translation>
<translation>athchóiriú</translation>
```

```
<translation>
  <value>leasú</value>
</translation>
<translation>
  <value>athchóiriú</value>
</translation>
```

And one more thing...

LEXiCAL
COMPUTING

And one more thing...

```
translation: leasú
    pos: n-masc
translation: athchóiriú
    pos: n-masc
    usage: formal
```

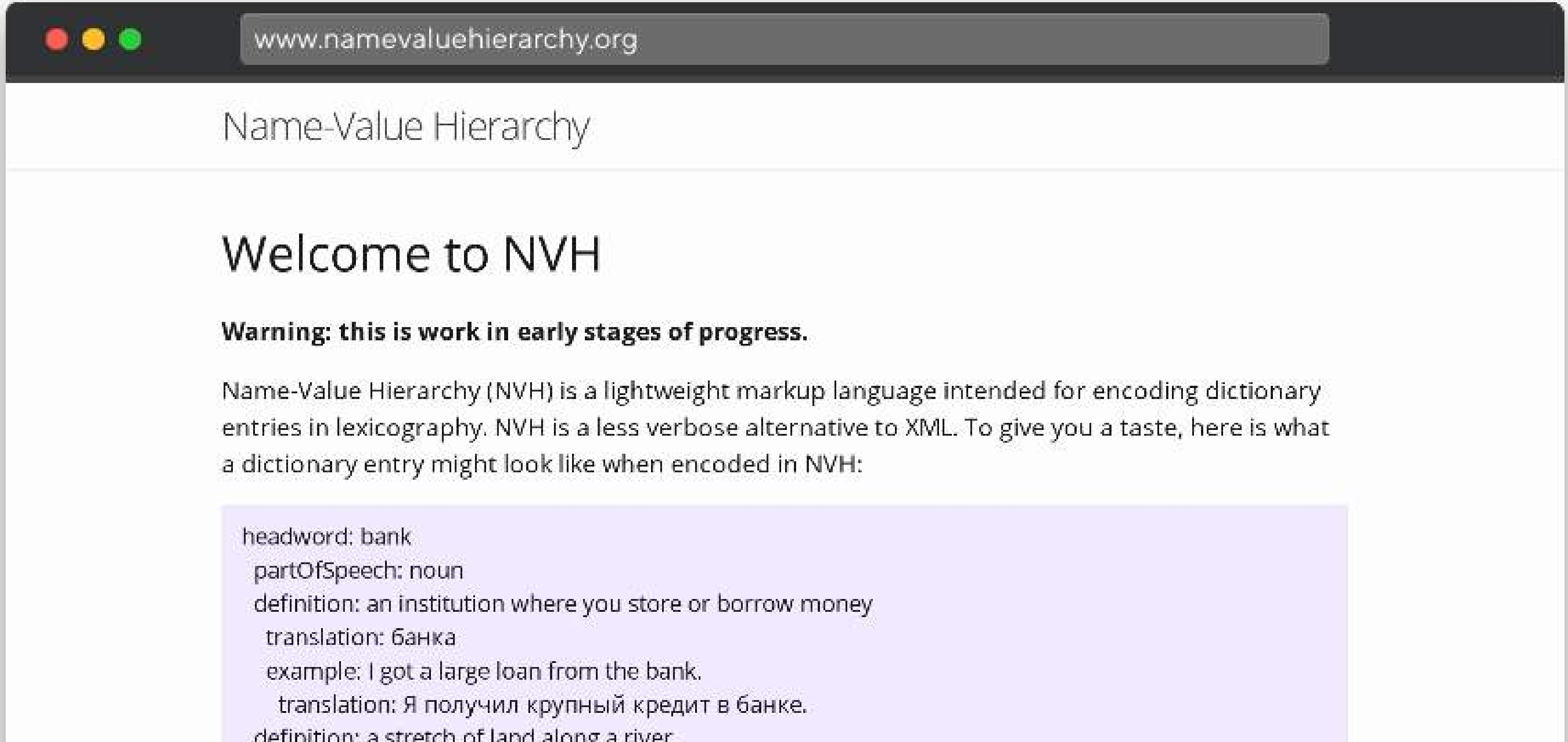
And one more thing...

```
translation: leasú
    pos: n-masc
translation: athchóiriú
    pos: n-masc
    usage: formal
```

Name-Value Hierarchy (NVH)

```
translation: leasú
    pos: n-masc
translation: athchóiriú
    pos: n-masc
    usage: formal
```

Name-Value Hierarchy (NVH)



Name-Value Hierarchy

Welcome to NVH

Warning: this is work in early stages of progress.

Name-Value Hierarchy (NVH) is a lightweight markup language intended for encoding dictionary entries in lexicography. NVH is a less verbose alternative to XML. To give you a taste, here is what a dictionary entry might look like when encoded in NVH:

```
headword: bank
  partOfSpeech: noun
  definition: an institution where you store or borrow money
    translation: банка
    example: I got a large loan from the bank.
      translation: Я получил крупный кредит в банке.
  definition: a stretch of land along a river
```

Name-Value Hierarchy (NVH)

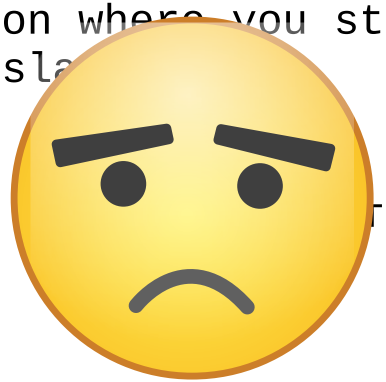
```
headword: bank
  partOfSpeech: noun
  definition: an institution where you store or borrow money
    translation: банка
    example: I got a large loan from the bank.
      translation: Я получил крупный кредит в банке.
  definition: a stretch of land along a river
    translation: берег
    example: The house is on the north bank of the river.
      translation: Дом находится на северном берегу реки.
```

NVH versus XML

```
<entry>
  <headword>bank</headword>
  <partOfSpeech>noun</partOfSpeech>
  <sense>
    <definition>an institution where you store or borrow money</definition>
    <translation>банка</translation>
    <exampleContainer>
      <example>I got a large loan from the bank.</example>
      <translation>Я получил крупный кредит в банке.</translation>
    </exampleContainer>
  </sense>
  <sense>
    <definition>a stretch of land along a river</definition>
    <translation>берег</translation>
    <exampleContainer>
      <example>The house is on the north bank of the river.</example>
      <translation>Дом находится на северном берегу реки.</translation>
    </exampleContainer>
  </sense>
```

NVH versus XML

```
<entry>
  <headword>bank</headword>
  <partOfSpeech>noun</partOfSpeech>
  <sense>
    <definition>an institution where you store or borrow money</definition>
    <translation>банка</transla
    <exampleContainer>
      <example>I got a large                    bank.</example>
      <translation>Я получил                    т в банке.</translation>
    </exampleContainer>
  </sense>
  <sense>
    <definition>a stretch of land along a river</definition>
    <translation>берег</translation>
    <exampleContainer>
      <example>The house is on the north bank of the river.</example>
      <translation>Дом находится на северном берегу реки.</translation>
    </exampleContainer>
  </sense>
```



Name-Value Hierarchy (NVH)

```
headword: bank
  partOfSpeech: noun
  definition: an institution where you store or borrow money
    translation: банка
    example: I got a large loan from the bank.
      translation: Я получил крупный кредит в банке.
  definition: a stretch of land along a river
    translation: берег
    example: The house is on the north bank of the river.
      translation: Дом находится на северном берегу реки.
```

www.namevaluehierarchy.org

Summary

Summary

Concepts:

- *purely structural markup*
- *matryoshkization*
- *headedness* (in lexicography, it’s everywhere!)
- *triples, tuples*

Support for headedness:

✗ XML

✗ JSON

✗ YAML

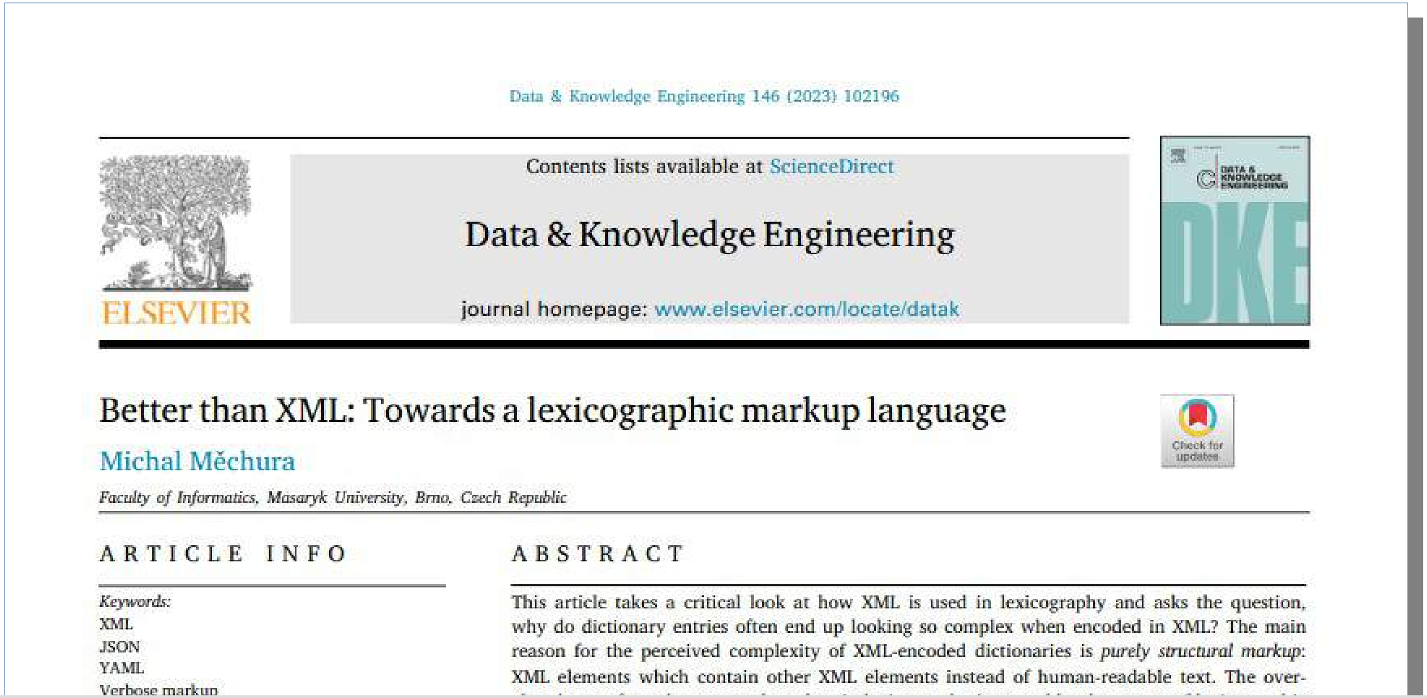
✓ SGML

✓ NVH

Thank you.

Michal Měchura
www.lexiconista.com
michmech@lexiconista.com

https://doi.org/10.1016/j.datak.2023.102196



Data & Knowledge Engineering 146 (2023) 102196

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Better than XML: Towards a lexicographic markup language

Michal Měchura

Faculty of Informatics, Masaryk University, Brno, Czech Republic

ARTICLE INFO

Keywords:
XML
JSON
YAML
Verbose markup

ABSTRACT

This article takes a critical look at how XML is used in lexicography and asks the question, why do dictionary entries often end up looking so complex when encoded in XML? The main reason for the perceived complexity of XML-encoded dictionaries is *purely structural markup*: XML elements which contain other XML elements instead of human-readable text. The over-

Thank you.

Michal Měchura
www.lexiconista.com
michmech@lexiconista.com