Skolemization in Unranked Logics

Lia Kurtanidze\textsuperscript{a*} and Mikheil Rukhaia\textsuperscript{b}

\textsuperscript{a}Faculty of Informatics, Georgian University
\textsuperscript{b}I. Vekua Institute of Applied Mathematics, I. Javakhishvili Tbilisi State University

(Received March 30, 2018; Revised June 8, 2018; Accepted June 12, 2018)

In this paper we study skolemization for unranked logics with classical first-order semantics. Skolemization is a transformation on first-order logic formulae, which removes all existential quantifiers from a formula. This technique is vital in proof theory and automated reasoning, especially for refutation based calculi, like resolution, tableaux, etc. Here we extend skolemization procedure to unranked formulae and prove that the procedure is sound and complete.

**Keywords:** Skolemization, Unranked logics; Proof complexity, Normal forms.

**AMS Subject Classification:** 68T15, 03F07, 03B70, 68T27, 03F05.

1. Introduction

Skolemization is a well known technique in refutational theorem proving to eliminate existential quantifiers. It is sometimes called an extension method, because it introduces new symbols in the signature of a formula. A very important feature of skolemization is that it loses logical equivalence, but preserves sat- or validity-equivalence.

Skolemization procedure is well studied for classical first-order logic (see e.g. \cite{1, 5, 20}), Constrained Logic (see e.g. \cite{2}), Intuitionistic Logic (see e.g. \cite{3, 4}), Fuzzy Logics (see e.g. \cite{6}), Lukasiewicz Logic (see e.g. \cite{7}), Probabilistic Logic (see e.g. \cite{27}) and the like.

The aim of this paper is to define different kinds of skolemization procedures for unranked logics and discuss the effects in the proof complexities of skolem normal forms.

Unranked languages are based on unranked alphabet, where function and/or predicate symbols do not have a fixed arity. Thus, in these languages, inside a term or a formula, it is possible to have several different occurrences of the same function/predicate symbol with different number of arguments. Additional strength to these languages are given by the notions of sequence variables and sequence functions. Sequence variables can be instantiated with finite sequences of terms, whereas sequence functions are interpreted as multi-valued functions. These constructs seems to be higher-order, but they have a precisely defined first-order semantics (see e.g. \cite{16, 18, 22}).

In recent years, the usefulness of sequence variables and unranked symbols has been illustrated in practical applications related to XML \cite{11–15, 25}, knowledge
representation [16, 18], automated reasoning [8, 19, 21, 23, 26], just to name a few. In this extent, it is well known that on the one hand, tableaux method is widely used for reasoning in Semantic Web and Knowledge Representation. On the other hand, skolemization procedure is crucial for the tableaux method. In [17] a tableaux calculus for unranked logics was presented, but the skolemization procedure itself was not defined. With the present paper we intend to fill the gap.

2. Preliminaries

We define first-order unranked predicate logic in the similar way as it was given in [17, 22]. We consider an alphabet $\mathcal{A}$ consisting of the following pairwise disjoint sets of symbols:

- the set $\mathcal{V}_i$ of individual variables, denoted by $x^i, y^i, z^i, \ldots$,
- the set $\mathcal{V}_s$ of sequence variables, denoted by $x, y, z, \ldots$,
- the set $\mathcal{F}_i^r$ of fixed arity (ranked) individual function symbols, denoted by $f^r, g^r, \ldots$,
- the set $\mathcal{F}_i^u$ of flexible arity (unranked) individual function symbols, denoted by $f^u, g^u, \ldots$,
- the set $\mathcal{F}_s^r$ of fixed arity (ranked) sequence function symbols, denoted by $\overrightarrow{f}, \overrightarrow{g}, \ldots$,
- the set $\mathcal{F}_s^u$ of flexible arity (unranked) sequence function symbols, denoted by $\overrightarrow{f}^u, \overrightarrow{g}^u, \ldots$,
- the set $\mathcal{P}_r$ of ranked predicate symbols, denoted by $p, q, \ldots$,
- the set $\mathcal{P}_u$ of unranked predicate symbols, denoted by $p^u, q^u, \ldots$,
- logical connectives and quantifiers $\neg, \land, \lor, \exists, \forall$,
- auxiliary symbols: parentheses and the comma.

Each ranked symbol has a unique arity (rank) associated with it. If the rank of a function symbol is 0, then it is called a constant. Unranked symbols do not have a fixed arity. Each set of variables, function symbols, and predicate symbols is countable infinite. The letter $\mathcal{V}$ denotes the set $\mathcal{V}_i \cup \mathcal{V}_s$ and its elements are denoted by $x, y, z, \ldots$.

The terms are defined as individual and sequence terms over $\mathcal{A}$ in the following inductive way:

\[
\begin{align*}
t &::= x^i | f^r(t_1, \ldots, t_n) | f^u(\overrightarrow{s}_1, \ldots, \overrightarrow{s}_m) & \text{individual terms} \\
\overrightarrow{s} &::= t | \overrightarrow{f}(t_1, \ldots, t_n) | \overrightarrow{f}^u(\overrightarrow{s}_1, \ldots, \overrightarrow{s}_m), n \geq 0 & \text{sequence terms}
\end{align*}
\]

An atom is a formula of the form $p(t_1, \ldots, t_n), n \geq 0$ and $p^u(\overrightarrow{s}_1, \ldots, \overrightarrow{s}_m), m \geq 0$, where $p \in \mathcal{P}_r$ is an $n$-ary predicate symbol, and $p^u \in \mathcal{P}_u$ is a flexible arity predicate symbol. Formulas are built as usual from atomic formulas and logical connectives $\neg, \land, \lor, \rightarrow, \exists$, and $\forall$. Quantifications are allowed on both, individual as well as on sequence variables. We use the letters $A, B, \ldots$ to denote formulas.

A substitution is a finite set of distinct variable bindings, where variable binding is either expression $x^i \mapsto t$ or $\overrightarrow{x} \mapsto \overrightarrow{s}$. Substitutions are denoted by $\sigma, \theta$ and the empty substitution is denoted by $\epsilon$. Application of a substitution $\sigma$ on term $t$ and formula $A$ is defined in the similar way as it is given in [24] and is denoted by $t\sigma$ and $A\sigma$, respectively.
The semantics of our language is assumed to be classical first-order semantics, defined in a similar way as it is given in [22, 24]. The notions of interpretation, evaluation, etc., are standard. Note that, as usual, \( \forall \) and \( \exists \) are dual to each other in our setting too; i.e., \( \forall^d \equiv \exists \) and \( \exists^d \equiv \forall \).

The following definitions provide some terminology, vital for the next section.

**Definition 2.1:** (Strong and weak quantifiers). The polarity of a quantifier is defined in the following inductive way:

- \((Qx)\) is positive in \((Qx)A\).
- if \((Qx)\) is positive (negative) in \(A\) or in \(B\), then it is positive (negative) in \(A \land B\) and in \(A \lor B\).
- if \((Qx)\) is positive (negative) in \(B\), then it is positive (negative) in \(A \rightarrow B\).
- if \((Qx)\) is positive (negative) in \(A\), then it is negative (positive) in \(\neg A\) and in \(A \rightarrow B\).

Positive universal (existential) quantifiers and negative existential (universal) quantifiers are called **strong** (weak) quantifiers.

**Definition 2.2:** (Quantifier scope). Let \(A\) be a formula and let \((Qx)B\) be its subformula. We say, that \(B\) is in the scope of \((Qx)\) and for every subformula \((Q'y)C\) of \(B\), the quantifier \((Q'y)\) is in the scope of the quantifier \((Qx)\).

Let \(A\) be a formula and \(x\) a variable occurring in \(A\). We say that an occurrence of \(x\) is bound in \(A\), if it occurs in the scope of the quantifier \((Qx)\), otherwise its occurrence is free. If all occurrences of \(x\) are bound in \(A\), then we say that \(x\) is a bound variable in \(A\). Please note, that it can be the case, that one variable has free and bound occurrences in a formula, but we can avoid such cases by renaming bound variables. In the similar way we avoid cases when one variable is bound by two or more quantifiers. We say that \(A\) is a closed formula, if it does not contain free variables.

**Definition 2.3:** (Quantifier omission). Let \(A\) be a formula. \(A_{-(Qx)}\) denotes the formula \(A\), where an occurrence of the quantifier \((Qx)\) is omitted.

**Remark 1** : Note that the bound variable \(x\) of \(A\) becomes free in \(A_{-(Qx)}\).

**Example 2.4** Let \(A\) be the formula \(B \land (\forall x^i)(\exists x^i)p^u(x^i, \overline{y})\). Then \(A_{-(\forall x^i)}\) is the formula \(B \land (\exists x^i)p^u(x^i, \overline{y})\) and \(A_{-(\exists x^i)}\) is \(B \land (\forall x^i)p^u(x^i, \overline{y})\).

We say, that a formula is in the prenex form, if it has the form \((Q_1x_1)\cdots(Q_nx_n)A\), where \(A\) is a quantifier-free formula. Every formula can be transformed into a prenex form using the following quantifier shifting rules:

\[
\neg(Qx)A \iff (Q^d x)\neg A \quad (\exists x)(A \lor B) \iff (\exists x)A \lor (\exists x)B \\
(\forall x)(A \land B) \iff (\forall x)A \land (\forall x)B \\
(\exists x)(A \rightarrow B) \iff (\forall x)A \rightarrow (\exists x)B
\]

\[\text{T}o \text{ transform a formula into prenex form, the quantifier shifting rules are used from right to left direction.} \]

\[\text{The other direction is used to transform a formula into antiprenex form.}\]
and if $x$ is not free in $B$, then

$$
(Qx)(A \lor B) \iff (Qx)A \lor B \\
(Qx)(A \land B) \iff (Qx)A \land B \\
(Qx)(A \rightarrow B) \iff (Q^d)xA \rightarrow B \\
(Qx)(B \rightarrow A) \iff B \rightarrow (Qx)A
$$

3. Skolemization

In a refutational calculus skolemization is a removal of the existential quantifiers from formulas. There are various ways to define skolemization:

**Prenex:** the traditional way to get skolem normal form of a formula. First, the formula is transformed to prenex normal form and then existential quantifiers are removed by replacing the corresponding bound variables by new $n$-ary function symbols, $n \geq 0$, where $n$ is the number of universal quantifiers, preceding the existential one.

**Structural:** this method does not need transformation to the prenex normal form. It is a bit more general, because it can eliminate strong quantifiers from a formula. The rule is similar – strong quantifier $(Qx)$ depends on the weak quantifiers having $(Qx)$ in their scope. It is possible to remove weak quantifiers in the same way, but it is called *Herbrandization* in the literature (see e.g. [9, 10]).

**Antiprenex:** this method is similar to structural skolemization, but contrary to the prenex normal form, quantifiers are shifted deep inside the formula using quantifier shifting rules, to minimize the range of quantifiers. This leads to smaller skolem terms in general.

Below we consider structural skolemization, because it is more general and does not need any preprocessing steps like prenex or antiprenex. We extend definitions from [1] to first-order unranked formulae.

**Definition 3.1:** (Structural skolemization). Let $A$ be a closed unranked formula, where no variable is bound by two quantifiers. We define a structural skolemization operator $sk()$ in the following way:

- If $A$ does not contain strong quantifiers, then $sk(A) = A$.
- Let $(Qx')$ be the first strong quantifier occurring in $A$. If $(Qx')$ is in the scope of weak quantifiers $(Q_1x_1), \ldots, (Qnx_n), n \geq 0$, then we distinguish the following cases:
  - if $n > 0$ and $x_j \in V_s$ for any $j = 1, \ldots, n$, then take a fresh flexible arity individual function symbol $f^{n}$ not occurring in $A$ and $sk(A) = sk(A_-(Qx'))[x' \mapsto f^{n}(x_1, \ldots, x_n)]$;
  - otherwise, take a fresh $n$-ary individual function symbol $f^n$ not occurring in $A$ and $sk(A) = sk(A_-(Qx'))[x' \mapsto f^n(x_1, \ldots, x_n)]$; note that, if $n = 0$, then $f^n$ is a constant.
- Let $(Q\overline{x})$ be the first strong quantifier occurring in $A$. If $(Q\overline{x})$ is in the scope of weak quantifiers $(Q_1x_1), \ldots, (Qnx_n), n \geq 0$, then we distinguish the following cases:
  - if $n > 0$ and $x_j \in V_s$ for any $j = 1, \ldots, n$, then take a fresh flexible arity sequence function symbol $\overline{f}^{n}$ not occurring in $A$ and $sk(A) = sk(A_-(Q\overline{x})[\overline{x} \mapsto \overline{f}^{n}(x_1, \ldots, x_n)]$

interesting is to consider either from the last two cases, e.g.:

The skolem terms will be similar to the one from $A$

\( \exists x^{i} (\forall y^{i}) p^{n}(x^{i}, y^{i}) \land (\forall \bar{x}) q^{n}(\bar{x}) \land (\exists \bar{y}) (p^{n}(\bar{y}) \lor q^{n}(\bar{y})) \)

and compute its structural, anti-prenex and prenex skolem forms. According to Definition 3.1, we get:

\[ \exists x^{i} \left( p^{u}(x^{i}, f^{r}(x^{i})) \land q^{u}(\bar{f}) \right) \land (\exists \bar{y}) (p^{u}(\bar{y}) \lor q^{u}(\bar{y})) \]

Note, that $f^{r}$ and $\bar{f}$ are different function symbols, individual and sequence function symbols, respectively, both with arity 1.

If we apply quantifier shifting rules, we can obtain the anti-prenex form of $A$:

\[ A' : (\exists x^{i})(\forall y^{i}) p^{n}(x^{i}, y^{i}) \land (\forall \bar{x}) q^{n}(\bar{x}) \land (\exists \bar{y}) (p^{n}(\bar{y}) \lor q^{n}(\bar{y})) \]

Again, by Definition 3.1, we get:

\[ (\exists x^{i}) p^{u}(x^{i}, f^{r}(x^{i})) \land q^{u}(\bar{f}) \land (\exists \bar{y}) (p^{u}(\bar{y}) \lor q^{u}(\bar{y})) \]

Note, that here $\bar{f}$ (with arity 0) is a constant and does not depend on $x^{i}$.

There are several prenex forms of $A$, namely, the order of quantifiers can be $(\exists x^{i})(\forall y^{i})(\forall \bar{x})(\exists \bar{y})$, $(\exists x^{i})(\forall \bar{x})(\forall y^{i})(\exists \bar{y})$, $(\exists \bar{y})(\exists x^{i})(\forall \bar{x})(\forall y^{i})$, or $(\exists \bar{y})(\exists x^{i})(\forall \bar{x})(\forall y^{i})$. The skolem terms will be similar to the one from $A$ in the first two cases, so interesting is to consider either from the last two cases, e.g.:

\[ A': (\exists \bar{y})(\exists x^{i})(\forall y^{i})(\forall \bar{x}) (p^{n}(x^{i}, y^{i}) \land q^{n}(\bar{x}) \land (p^{n}(\bar{y}) \lor q^{n}(\bar{y}))) \]
According to Definition 3.1, we get:

\[(\exists y)(\exists x) (p^u(x, f^u(y, x')) \land q^u(\overline{f}^u(y, x')) \land (p^u(y) \lor q^u(y)))\]

where, again, \(f^u\) and \(\overline{f}^u\) are different function symbols, individual and sequence function symbols, respectively, both with the flexible arity.

**Theorem 3.4:** (Soundness). If \(A\) is valid, then \(sk(A)\) is also valid.

**Proof:** Proceed by induction on the number of strong quantifiers in \(A\). The base case is trivial: \(A = sk(A)\).

Assume, that the theorem holds for all formulae with at most \(n\) strong quantifiers and \(A\) contains \(n + 1\) strong quantifiers. Then we select the first strong quantifier \((Qx)\), and assume it is in the scope of weak quantifiers \((Q_1x_1), \ldots, (Q_mx_m), m \geq 0\) in \(A\). We shift those quantifiers to the front using quantifier shifting rules, then we get an equivalent formula of \(A\), \((\exists x_1)\ldots(\exists x_m)(\forall x)B\), where \(B = A^{-}(Qx)(Q_1x_1)\ldots(Q_mx_m)\).

By Definition 3.1, \(sk((\exists x_1)\ldots(\exists x_m)(\forall x)B) = sk((\exists x_1)\ldots(\exists x_m)B\sigma)\), where \(\sigma\) is a substitution built according to Definition 3.1 (case distinction on \(A\) in \(A\), not relevant for this proof).

By the induction hypothesis, if \((\exists x_1)\ldots(\exists x_m)B\sigma\) is valid, then \(sk((\exists x_1)\ldots(\exists x_m)B\sigma)\) is valid too. Thus, it is sufficient to show that \((\exists x_1)\ldots(\exists x_m)(\forall x)B \rightarrow (\exists x_1)\ldots(\exists x_m)B\sigma\), which is trivially valid. \(\square\)

**Theorem 3.5:** (Completeness). If \(sk(A)\) is valid, then \(A\) is also valid.

**Proof:** Proceed by induction on the number of strong quantifiers in \(A\). The base case is trivial: \(sk(A) = A\).

Assume, that the theorem holds for all formulae with at most \(n\) strong quantifiers and \(A\) contains \(n + 1\) strong quantifiers. Then we select the first strong quantifier \((Qx)\) and distinguish the following cases:

- \((Qx)\) is not in the scope of weak quantifiers in \(A\). Then, by Definition 3.1 \(sk(A) = sk(A_{-(Qx)}\sigma)\), where \(\sigma\) is either \([x \mapsto f']\) (if \(x \in V_i\)) or \([x \mapsto \overline{f}']\) (if \(x \in V_s\)), for \(f'\) (or \(\overline{f}'\)), with arity 0, not occurring in \(A\). By the induction hypothesis, if \(sk(A_{-(Qx)}\sigma)\) is valid, then \(A_{-(Qx)}\sigma\) is valid too. Thus, it is sufficient to show that if \(A_{-(Qx)}\sigma\) is valid, then \(A\) is also valid.

  Note that \(A\) is equivalent to \(A'\): \((\forall x)A_{-(Qx)}\) (using quantifier shifting rules \((Qx)\) is shifted to front) and assume that \(A'\) (and thus \(A\)) is not valid. Then there is an interpretation \(\Gamma\), that falsifies \(A'\). By the standard definition of interpretations, there is also an interpretation \(\Gamma'\) that differs from \(\Gamma\) in the interpretation of \(x\) and falsifies \(A'\). So \(f'\) (or \(\overline{f}'\)) does not occur in \(A\), we can extend \(\Gamma\) to \(\Gamma''\), where \(x\) is interpreted as \(f'\) (or \(\overline{f}'\)). Then \(\Gamma''\) falsifies \(A_{-(Qx)}\sigma\), implying that \(A_{-(Qx)}\sigma\) is not valid either.

- \((Qx)\) is in the scope of weak quantifiers \((Q_1x_1), \ldots, (Q_mx_m), m > 0\) in \(A\). We shift those quantifiers in front using quantifier shifting rules, then we get an equivalent formula of \(A\), \(A': (\exists x_1)\ldots(\exists x_m)(\forall x)B\), where \(B = A_{-(Qx)(Q_1x_1)\ldots(Q_mx_m)}\). By Definition 3.1, \(sk(A') = sk((\exists x_1)\ldots(\exists x_m)B\sigma)\), where \(\sigma = [x \mapsto f(x_1, \ldots, x_m)]\),

\(^{1}\)Note that weak quantifiers become existential ones and strong quantifiers become universal ones when shifted in front.
and $f$ is either in $\mathcal{F}_i^r$, $\mathcal{F}_i^s$, $\mathcal{F}_s^r$ or $\mathcal{F}_s^s$ depending on $x, x_1, \ldots, x_m$. Without loss of generality we assume that $f \in \mathcal{F}_s^s$, i.e., $x \in \mathcal{V}_s$, and there is $k \in \{1, \ldots, m\}$, s.t., $x_k \in \mathcal{V}_s$; the other cases are similar.

By the induction hypothesis, if $sk((\exists x_1)\ldots(\exists x_m)B\sigma)$ is valid, then $(\exists x_1)\ldots(\exists x_m)B\sigma$ is valid too. Thus, it is sufficient to show that if $(\exists x_1)\ldots(\exists x_m)B\sigma$ is valid, then $A'$ (and thus $A$) is also valid.

Again, assume that $A'$ is not valid, then there is an interpretation $\Gamma$ that falsifies $A'$ and is a model of $\neg A'$: $(\forall x_1)\ldots(\forall x_m)(\exists x)\neg B$. By axiom of choice, there is a function $\overline{y}^u$ of $x_1, \ldots, x_m$ to assign to $x$, and since $f$ does not occur in $A$, we can take $\overline{y}^u = f$ and construct an interpretation $\Gamma'$, extension of $\Gamma$, that will be a model of $(\forall x_1)\ldots(\forall x_m)\neg B\sigma$, thus falsify $(\exists x_1)\ldots(\exists x_m)B\sigma$. This proves the theorem.

\[ \square \]

**Corollary 3.6:** $A$ is sat-equivalent to $\neg sk(\neg A)$, i.e., $A$ is satisfiable iff $\neg sk(\neg A)$ is.

**Proof:** Direct consequence of theorems 3.4 and 3.5. \[ \square \]

It is easy to see, that different skolemization methods produce formulae of the similar length (the number of symbols) and logical complexity (the number of logical connectives). In [5] it was shown, that in terms of proof complexity, the particular form of skolemization actually matters, since it might destroy some information encoded inside a formula. The complexity analysis is based on the notion of Herbrand complexity – the minimum size of Herbrand disjunction of a formula. The detailed discussion can be found in [1, 5]. The complexity results obtained in [1, 5] can be extended to unranked logics in a straightforward way.

To summarize the mentioned results, as it was stated in [1], if a given formula is already in the prenex form, then “antiprenex skolemization may give a non-elementary speed-up over the structural one” during the proof search. Thus, skolemization should be considered as “an integral part of the inference process and not as a preprocessing step of minor importance”.

### 4. Conclusion

We have investigated skolemization procedure for first-order unranked formulas. We defined the algorithm and proved that it is sound and complete. Although we did not give a formal proof, it is easy to see, that the complexity relationship between different kind of skolemization procedures hold for unranked formulas as well.

**Acknowledgements**

This work was supported by the Shota Rustaveli National Science Foundation project PhD_F17_218.

**References**

Reasoning (KR), 2014, 1-10
In Proceedings of the 14th International Conference on Principles of Knowledge Representation and
Skolemization for weighted first-order model counting