**Калегин С.Н.**

ЗАО «МНИТИ», Москва, РФ

# ВАЖНОСТЬ ВЫБОРА ОСНОВНОГО ИДЕНТИФИКАЦИОННОГО ПРИНЦИПА ПРИ ПРОЕКТИРОВАНИИ ЯЗЫКОВЫХ ОПРЕДЕЛИТЕЛЕЙ

**АННОТАЦИЯ**

*В статье приводятся результаты сравнительного исследования различных программ определения языковой принадлежности текста, в основу алгоритмов которых заложены различные способы идентификации, с целью выявления зависимости их характерных особенностей от выбранных идентификационных принципов. Приведённые материалы наглядно демонстрируют причины общности, преимуществ и недостатков рассмотренных решений проблемы определения языковой принадлежности информации.*

**КЛЮЧЕВЫЕ СЛОВА**

*Идентификационный принцип, способ языковой идентификации; языковая идентификация; идентификация языка; определение языка; определение языковой принадлежности.*

**Kalegin S.N.**

CJSC MNITI, Moscow, RF

# THE IMPORTANCE OF CHOOSING THE MAIN IDENTIFICATION PRINCIPLE IN THE DESIGN OF LANGUAGE IDENTIFIER

**ABSTRACT**

*The article presents the results of a comparative research of different types of texts language identification programs, based on different methods of identification, to identify the dependence of the characteristics of the selected identification principles. The given materials clearly demonstrate the reasons of community advantages and disadvantages of the examined solutions to the problem of language detection.*

**KEYWORDS**

*Identification principle; identification method; language identification; language detection.*

В данной статье приводятся результаты сравнительного исследования различных программ определения языковой принадлежности текста, в основу алгоритмов которых заложены различные способы идентификации, что позволило выявить существенные преимущества и недостатки как самих способов, так и их программной реализации [1]. Этот обзор позволил выявить зависимость эффективности работы языковых определителей от идентификационных принципов и способов, заложенных в их основу. Результаты проведённого анализа помогут программистам улучшить существующие и сделать более совершенными разрабатываемые программы, а пользователям более осознанно подходить к их выбору в зависимости от типа применения.

На сегодняшний день существует множество языковых определителей, однако каждый из них обладает некоторыми недостатками, обусловленными характерными особенностями используемых способов идентификации, показанных на рисунке 1.

Как видно из приведённой блок-схемы [2], существуют различные способы и подходы языковой идентификации неструктурированного текста, а в качестве определяющих идентифицирующих элементов могут быть выбраны различные компоненты: символы, их сочетания или наборы – грамм, служебные слова и частицы, значимые слова и т.д., от выбора и состава которых зависит эффективность и результативность процесса определения языка. Таким образом, выбор базового принципа, способа и типа идентифицирующих элементов обуславливает характерные особенности программ, в основу которых они положены, и существенно влияет на результаты их применения, что будет показано ниже.