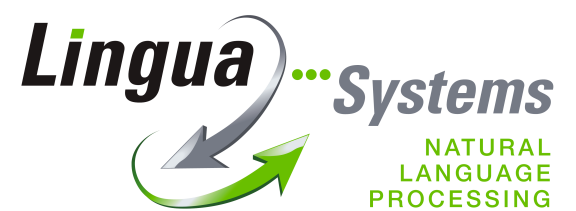


Software-Spezifikation

für

lidc

Beschreibt Version 1.3.0



1 Einführung

Dies ist die Software-Spezifikation für **lidc** Version 1.3.0.

lidc ist eine Anwendung, mit der sich sowohl die Sprache als auch die Zeichenkodierung einer Text-, HTML-, XML- oder E-Mail- Eingabe bestimmen lässt.

Die Spezifikation liefert eine allgemeine Beschreibung der Software und eine detaillierte Beschreibung des Leistungsumfangs.

2 Allgemeine Beschreibung der Software

lidc ist eine Kommandozeilenanwendung für Unix-artige Betriebssysteme, die Sprache und Zeichenkodierung einer Eingabe bestimmt. Es werden verschiedene, weit verbreitete, Eingabeformate unterstützt. Die Ausgabe der Ergebnisse kann durch den Benutzer sehr frei bestimmt werden.

2.1 Benutzermerkmale

Generell kann die Software von jedem Benutzer verwendet werden, der über Grundkenntnisse in der Verwendung der Kommandozeile unter Unix-artigen Betriebssystemen, wie z.B. Linux oder Solaris, verfügt.

Es handelt sich bei **lidc** um eine Expertensoftware, die in jedem Umfeld gewinnbringend zum Einsatz gebracht werden kann, bei dem die Erkennung von Sprache oder Zeichenkodierung eines Textes erforderlich oder hilfreich ist.

2.2 Betriebsumgebung

lidc ist eine Anwendung, die zum Einsatz unter Unix-artigen Betriebssystemen konzipiert ist. Eine Standardversion der Software kann für die folgenden Betriebssysteme bezogen werden:

- Debian GNU/Linux (x86): *Lenny*
- Ubuntu GNU/Linux (x86): *LTS (10.04)*
- Solaris (Sparc): *10*
- FreeBSD (x86): *7, 8*
- Mac OS X (x86): *Tiger, Leopard, Snow Leopard*

Die korrekte Funktionalität wird unter anderen Distributionen und/oder Versionen eines Betriebssystems zu großer Wahrscheinlichkeit ebenfalls gegeben sein, ist allerdings nicht Umfang der Gewährleistung.

Versionen für weitere Betriebssysteme und/oder Distributionen können auf Anfrage ebenfalls bereitgestellt werden.

2.3 Abhängigkeiten

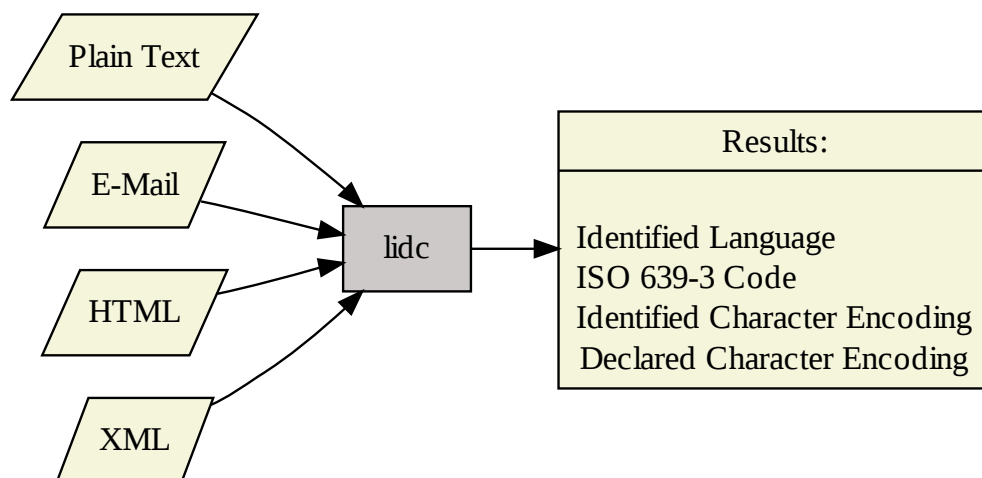
`lidc` weist außer der Standard C- und Thread-Bibliothek der jeweiligen Systeme keine weiteren Abhängigkeiten auf.

2.4 Ressourcenverbrauch

Der Bedarf an Arbeitsspeicher ist stark abhängig von der Größe und dem Typ der Eingabe-Datei. Mindestens werden jedoch 30 KiB benötigt.

3 Leistungsumfang der Software

Die Kommandozeilenanwendung `lidc` bestimmt in der hier beschriebenen Version 28 Sprachen und 35 Zeichenkodierungen. Zusätzlich werden 10 Sprachen in transliterierter Form unterstützt. Die Eingabe kann in verschiedenen Formaten übergeben werden. Zurückgegeben werden die erkannte Sprache und ihr dreistelliger ISO 639-3 Code, die erkannte Zeichenkodierung sowie - falls vorhanden - die angegebene Zeichenkodierung der Eingabe.



Im Folgenden werden die unterstützten Eingabeformate, Sprachen, Kodierungen und die Ausgabe näher bestimmt. Neben dieser funktionellen Beschreibung werden Sicherheitsaspekte, Qualitätsmerkmale und Einschränkungen der Software aufgeführt.

3.1 Benutzerschnittstelle

Die gesamte Funktionalität wird über eine Kommandozeilenanwendung bereitgestellt. Eingabe und Ausgabe werden über entsprechende Parameter angepasst. Alle Parameter werden im Benutzerhandbuch zur Software aufgeführt.

3.2 Eingabe

3.2.1 Unterstützte Eingabeformate

`lidc` beherrscht durch eine eingebaute Filtertechnik die Verarbeitung der folgenden Formate:

1. Plain Text (MIME-Type: `text/plain`)
2. HTML: HTML (alle Versionen), X-HTML
3. XML
4. E-Mail (RFC 822)
5. E-Mail: `text/plain`, `text/html`, `multipart/mixed`, `multipart/alternative`, `multipart/digest`, `message/rfc822`, `multipart/parallel` (RFC 2045-2049: MIME)
6. E-Mail: `multipart/related` (RFC 2387)
7. E-Mail: `multipart/report` (RFC 3462)
8. E-Mail: `multipart/signed` (RFC 1847)

Das jeweilige Format kann durch ein Kommandozeilenargument ausgewählt oder automatisch durch eine Auswertung der Datei-Endung bestimmt werden.

Die Eingabe kann direkt als Datei oder durch eine Verkettung an `lidc` übergeben werden, so dass die Software in Pipes eingesetzt werden kann.

`lidc` beherrscht auch die Verarbeitung von UTF-16 und UTF-32 kodierten Dateien (sowohl im Big- als auch Little Endian Format). Diese Kodierungen sind allerdings nicht für E-Mails unterstützt.

3.2.2 Unterstützte Sprachen und Kodierungen

Es können derzeit 28 Sprachen erkannt werden. Die unterstützten Zeichenkodierungen umfassen sowohl gängige als auch ältere Zeichenkodierungen für die jeweiligen Sprachen. Die einer UTF-16 oder UTF-32 kodierten Eingabe zugrundeliegende Bytereihenfolge wird ebenfalls bestimmt und durch „UTF-16BE“, „UTF-16LE“, „UTF-32BE“ und „UTF-32LE“ mit angegeben.

Sprache	ISO 639-3 Code	Zeichenkodierungen
Bokmål (Norwegisch)	nob	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII
Bulgarisch	bul	UTF-32, UTF-16, UTF-8, ISO-8859-5, Windows-1251, MacCyrillic, CP 855, CP 866, KOI8-R
Dänisch	dan	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII
Deutsch	deu	UTF-32, UTF-16, UTF-8, ISO-8859-1, ISO-8859-15, Windows-1252, MacRoman, CP 850, ASCII

Sprache	ISO 639-3 Code	Zeichenkodierungen
Englisch	eng	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII
Estnisch	est	UTF-32, UTF-16, UTF-8, ISO-8859-4, Windows-1257, MacCentralEurope, CP 775, ASCII
Finnisch	fin	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII, ISO-8859-15
Französisch	fra	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII, ISO-8859-15
Griechisch	ell	UTF-32, UTF-16, UTF-8, ISO-8859-7, Windows-1253, MacGreek, CP 737
Irish (Gälisch)	gle	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII
Italienisch	ita	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII, ISO-8859-16
Lettisch	lav	UTF-32, UTF-16, UTF-8, ISO-8859-4, Windows-1257, MacCentralEurope, CP 775, ASCII
Litauisch	lit	UTF-32, UTF-16, UTF-8, ISO-8859-4, Windows-1257, MacCentralEurope, CP 775, ASCII
Maltesisch	mlt	UTF-32, UTF-16, UTF-8, ISO-8859-3
Mandarin (Chinesisch)	cmn	UTF-32, UTF-16, UTF-8, Big5, GB2312
Nynorsk (Norwegisch)	nno	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII
Niederländisch	nld	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII, ISO-8859-15
Polnisch	pol	UTF-32, UTF-16, UTF-8, ISO-8859-2, Windows-1250, MacCentralEurope, CP 852, ISO-8859-16
Portugiesisch	por	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII, ISO-8859-15
Rumänisch	ron	UTF-32, UTF-16, UTF-8, ISO-8859-2, Windows-1250, MacRomanian, CP 852
Russisch	rus	UTF-32, UTF-16, UTF-8, ISO-8859-5, Windows-1251, MacCyrillic, CP 855, CP 866, KOI8-R
Schwedisch	swe	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII
Slowakisch	slk	UTF-32, UTF-16, UTF-8, ISO-8859-2, Windows-1250, MacCentralEurope, CP 852
Slowenisch	slv	UTF-32, UTF-16, UTF-8, ISO-8859-2, Windows-1250, MacCentralEurope, CP 852, ASCII, ISO-8859-16
Spanisch	spa	UTF-32, UTF-16, UTF-8, ISO-8859-1, Windows-1252, MacRoman, CP 850, ASCII, ISO-8859-15
Tschechisch	ces	UTF-32, UTF-16, UTF-8, ISO-8859-2, Windows-1250, MacCentralEurope, CP 852
Ukrainisch	ukr	UTF-32, UTF-16, UTF-8, Windows-1251, MacUkrainian, KOI8-U
Ungarisch	hun	UTF-32, UTF-16, UTF-8, ISO-8859-2, Windows-1250, MacCentralEurope, CP 852, ISO-8859-16

Darüber hinaus werden auch transliterierte Texte der folgenden Sprachen in den in angegebenen Transliterationen erkannt:

Sprache	Transliteration	Zeichenkodierungen
Bulgarisch	ISO 9	UTF-32, UTF-16, UTF-8, ASCII
	DIN 1460	UTF-32, UTF-16, UTF-8, ASCII, Windows-1250
	Streamlined System	UTF-32, UTF-16, UTF-8, ASCII, Windows-1250
Deutsch	gebräuchlich	UTF-32, UTF-16, UTF-8, ASCII, ISO-8859-1
Griechisch	ISO 843	UTF-32, UTF-16, UTF-8, ASCII
	DIN 31634	UTF-32, UTF-16, UTF-8, ASCII
	Greeklish	UTF-32, UTF-16, UTF-8, ASCII, ISO-8859-1
Polnisch	gebräuchlich	UTF-32, UTF-16, UTF-8, ASCII, ISO-8859-1
Rumänisch	gebräuchlich	UTF-32, UTF-16, UTF-8, ASCII, ISO-8859-1
Russisch	ISO 9	UTF-32, UTF-16, UTF-8
	DIN 1460	UTF-32, UTF-16, UTF-8
Slowakisch	gebräuchlich	UTF-32, UTF-16, UTF-8, ASCII, ISO-8859-1
Slowenisch	gebräuchlich	UTF-32, UTF-16, UTF-8, ASCII, ISO-8859-1
Tschechisch	gebräuchlich	UTF-32, UTF-16, UTF-8, ASCII, ISO-8859-1
Ukrainisch	ISO 9	UTF-32, UTF-16, UTF-8
	DIN 1460	UTF-32, UTF-16, UTF-8

„gebräuchlich“ denotiert hierbei die üblicherweise angewendeten, nicht standardisierten Transliterationen, die dennoch häufig zur Anwendung kommen.

3.3 Ausgabe

3.3.1 Ergebnisse

Die folgenden Werte können als Ergebnisse zurückgeliefert werden:

- Die erkannte Sprache der Eingabe
Die Sprache wird nach ihrer englischen Bezeichnung ausgegeben. Zusätzlich kann der dreistellige ISO 639-3 Code ausgegeben werden.
- Die erkannte Zeichenkodierung
Die Zeichenkodierung wird entsprechend der Liste aus Kapitel 3.2.2 ausgegeben.
- Die angegebene Zeichenkodierung (falls vorhanden)
Wenn in einer Datei die Zeichenkodierung mit angegeben ist, wird diese zusätzlich als Ergebnis zurückgeliefert. Das Format entspricht dem der erkannten Zeichenkodierung, allerdings ausschließlich in Kleinbuchstaben.
- Der Name der Eingabe

3.3.2 Formatierung

`lidc` sieht die Ausgabe der Ergebnisse auf der Kommandozeile (`stdout`) vor. Das Format der Ausgabe und die darin enthaltenen Daten können vom Benutzer durch einen Formatstring angegeben werden.

Zum Referenzieren der Ergebnisse dienen Platzhalter für die Eingabe-Datei (%f), deren ermittelte (%e) und (gegebenenfalls) angegebene Zeichenkodierung (%d) und die bestimmte Sprache (%l) mit ihrem ISO 639-3 Code (%i). Darüber hinaus können, wie bei jedem anderen Formatstring auch, beliebige weitere Zeichen angegeben werden, so dass beispielsweise auch eine Ausgabe der Ergebnisse im CSV-, HTML- oder XML-Format möglich ist.

3.4 Fehlerbehandlung

`lidc` gibt beim Auftreten eines Fehlers eine entsprechende, englischsprachige Fehlermeldung aus und beendet sich mit einem Fehlercode.

3.5 Sicherheitsaspekte

`lidc` wurde auch im Hinblick auf Sicherheitsaspekte entwickelt und weist keine bekannten Sicherheitsmängel auf. Darüber hinaus legt die Software keine temporären Dateien im Dateisystem an und wertet keine Umgebungsvariablen aus.

3.6 Qualitätsmerkmale

`lidc` arbeitet sowohl schnell als auch zuverlässig und kann mit vielen weit verbreiteten Eingabeformaten umgehen. Die Geschwindigkeit der Verarbeitung ist abhängig von der zu Grunde liegenden Hardware und der Auslastung des Systems.

3.7 Einschränkungen

Eine 100%ige Erkennung von Sprache und Zeichenkodierung kann nach dem gegenwärtigen Stand der Technik auch von `lidc` nicht in jedem Fall uneingeschränkt garantiert werden. Es kann somit bei einigen Dokumenten vorkommen, dass eine falsche Sprache und/oder Zeichenkodierung erkannt wird.

Eine Eingabe kann nur dann sicher verarbeitet werden, wenn sie in einem der unterstützten Formate vorliegt (siehe Kapitel 3.2.1) oder im Vorfeld in ein solches konvertiert wurde.

Es können nur unterstützte Sprachen und Zeichenkodierungen (siehe Kapitel 3.2.2) erkannt werden.

Sprachen und Kodierungen, die nicht unterstützt sind, werden trotzdem der Sprache und/oder Kodierung zugewiesen, der sie am ähnlichsten sind.

Ebenso können keine mehrsprachigen Dokumente verarbeitet werden, ausgenommen Texte mit kleineren fremdsprachigen Einschüben.

Eine zu erkennende Eingabe muss eine bestimmte Mindestlänge aufweisen. Es sollten mindestens 25 Zeichen in verschiedenen Wörtern vorliegen.

Bei der Verarbeitung von *E-Mails* wird bei einer Multipart E-Mail (*MIME*) nur die erste Textnachricht ausgewertet.

Die Software liegt nur in englischer Sprache vor.

Das Handbuch ist sowohl auf Deutsch als auch auf Englisch verfügbar. Es wird ausschließlich in Form einer PDF Datei bereitgestellt.

4 Auslieferung

Eine vollständige Installation der Software umfasst die Kommandozeilenapplikation, das Benutzerhandbuch in deutscher und englischer Fassung und Manual-Seiten (nur in Englisch). Die Software belegt –je nach System– etwa 2 MiB Festplattenspeicher.

5 Referenzen

- Lingua-Systems' [lidc](http://www.lingua-systems.de/language-identifizier/lidc-application/) Produktwebseite, <http://www.lingua-systems.de/language-identifizier/lidc-application/>
- E-Mail betreffend:
 - RFC 2045, 2046, 2047, 2049 - „Multipurpose Internet Mail Extensions“
 - RFC 822 - „Standard for the Format of ARPA Internet Text Messages“
 - RFC 2387 - „The MIME Multipart/Related Content-type“
 - RFC 1847 - „Security Multiparts for MIME: Multipart/Signed and Multipart/Encrypted“
 - RFC 3462 - „The Multipart/Report Content Type for the Reporting of Mail System Administrative Messages“
- HTML betreffend:
 - W3C XHTML 1.0 Spezifikation: „The Extensible HyperText Markup Language“
 - W3C HTML 4.01 Spezifikation
- XML betreffend:
 - W3C Extensible Markup Language (XML) 1.0 Spezifikation
- Sprachkennungen : ISO 639-3:2007 - „Codes for the representation of names of languages“