

The MYLL Programming Language: C++ Resyntaxed and Extended

Eine Weiterentwicklung der C Style Syntax

zur Erlangung des Grades

Master of Science - Informatik

an der

Hochschule Niederrhein

Fachbereich Elektrotechnik und Informatik

vorgelegt von

Jan Helge Reitz

Geboren 1978-10-06 in Moers, Deutschland

Matrikelnummer: 756470

Datum: 2020-07-03

Prüfer: Prof. Dr. Jochen Rethmann

Zweitprüfer: Prof. Dr. Steffen Goebbels

Zusammenfassung

Ich schlage eine neue Syntax für C++ vor, welche den Umgang mit C++ unproblematischer und den Einstieg erleichtern soll.

Die von mir vorgeschlagene neue Syntax für C++ hat den Codenamen **Myll** (*frei von My Language*). Der Fokus des Sprach-Designs balanciert zwischen zwei grundlegenden Zielen; zum einen soll er alteingesessene C++ Programmierer abholen und dessen Arbeit in Zukunft leichter gestalten und zum anderen soll es Neulingen den Einstieg, in eine immer komplizierter werdende Sprache, erleichtern.

Um dieses Ziel zu erreichen habe ich viele Sprachkonstrukte auf die Probe gestellt und mir überlegt wie man diese besser gestalten kann, teils von anderen Sprachen inspiriert, aber auch komplett selbst erdacht.

Des Weiteren soll Myll einfach ergänzbar sein: Beispielsweise nutzen QT, UE4 und viele andere Bibliotheken *Makros*, welche an den richtigen Stellen in den Code eingesetzt werden und teilweise durch einen zusätzlichen Präprozessor ausgeführt werden. Diese verstreuten Makros können in Myll etwa durch ein Klassen- oder Variablen-Attribut ersetzt werden.

Umgesetzt ist Myll als ein Language to Language Compiler, auch Transpiler genannt. Dieser erzeugt als Ausgabe keinen Maschinencode, sondern Quellcode einer anderen Programmiersprache, in diesem Fall C++.

Dies soll ein *hackbarer* Transpiler sein; einfach herunterladen und die eigene Sprache erfinden!

Abstract

I propose a new syntax for C++, which should make the general usage of C++ less problematic and also make it easier to get started.

The new syntax for C++ proposed by me has the code name **Myll** (*free from My Language*). The focus of the language design balances between two basic goals; on the one hand, it should pick up established C++ programmers and make their work easier in the future, and on the other hand, it should make it easier for newcomers to get started with a language which has become more and more complicated.

In order to achieve this goal, I have put many language constructs to the test and thought about how they could be better designed, partly inspired by other languages, but also completely conceived by myself.

Furthermore Myll should be easy to extend: For example QT, UE4 and many other libraries use macros which are inserted at the right places in the code and are partially executed by an additional preprocessor. These scattered macros can be replaced by an attribute on a type or variable in Myll.

Myll is implemented as a language to language compiler, also called a transpiler. This compiler does not produce machine code as output, instead it produces source code of another programming language, in this case C++.

This Transpiler is meant to be *hackable*; download it and invent your own language!

Inhaltsverzeichnis

Zusammenfassung.....	2
Abstract.....	2
Inhaltsverzeichnis.....	3
1 Einleitung.....	6
1.1 Definitionen.....	6
1.1.1 Formatierung dieses Dokumentes.....	6
1.1.2 Begriffsdefinitionen.....	6
1.2 Idee / Motivation.....	9
1.2.1 Alteingesessene C++ Programmierer abholen.....	9
1.2.2 Neulingen den Einstieg erleichtern.....	9
1.3 Ziel.....	10
1.4 Grundlagen.....	10
1.4.1 Was macht C++ besonders?.....	10
1.4.2 Wieso noch eine neue Programmiersprache?.....	11
2 Analyse.....	12
2.1 Die Besonderheiten von C++.....	12
2.2 Vergleich von C++ mit anderen Sprachen.....	13
2.2.1 Vergleich mit C.....	13
2.2.2 Vergleich mit C# und Java.....	13
2.2.3 Vergleich mit D.....	14
2.2.4 Vergleich mit Rust.....	14
2.2.5 Vergleich mit Ruby.....	14
2.2.6 Vergleich von JavaScript mit TypeScript.....	15
2.3 Problematische Aspekte von C++.....	16
3 Konzept.....	17
3.1 Einleitung.....	17
3.1.1 Grundsätze der Sprache Myll.....	17
3.1.2 Fundamental Principles of Myll.....	17
3.1.3 Inspiration.....	17
3.1.4 Lerneffekte der verglichenen Sprachen.....	18
3.2 Design der Sprache Myll.....	19
3.2.1 Neue Keywords.....	19
3.2.2 Übersicht der Syntax von Myll.....	19
3.2.3 Gesamtbild.....	20
3.3 Lösung der problematischen Aspekte von C++.....	21
3.3.1 Deklarationen, Definitionen und Initialisierung.....	21
3.3.1.1 Reihenfolge und Prototypen.....	21
3.3.1.2 Aufteilung auf verschiedene Orte.....	21

3.3.1.3 Aufteilung im Kontext sich stetig verändernder Codebases.....	23
3.3.2 Fehleranfällige Variablen & Typen Deklarationen.....	23
3.3.2.1 Ambiguität einer Deklaration.....	23
3.3.2.2 Pointer, Referenz & Array.....	24
3.3.2.3 Pointer auf Arrays (Arithmetik auf Pointern oder auch nicht).....	25
3.3.2.4 Die Kombination von Pointern und Arrays.....	26
3.3.2.5 Funktionspointer.....	26
3.3.3 Schlechtes Standardverhalten.....	27
3.3.3.1 Ein Parameter Konstruktor.....	27
3.3.3.2 Private Ableitung.....	29
3.3.3.3 Alles kann Exceptions werfen.....	29
3.3.3.4 Switch fällt von selbst.....	30
3.3.4 Inkonsistente Syntax.....	31
3.3.4.1 Funktions- & Methodendeklarationen.....	31
3.3.4.2 Notwendigkeit des Semikolons.....	32
3.3.4.3 Verteilung der Attribute.....	32
3.3.4.4 Implizite Konvertierungen.....	33
3.3.5 Problematische Reihenfolge der Syntax.....	34
3.3.5.1 Unerwartete Reihenfolge der Operatoren.....	35
3.3.5.2 Pre- und Suffix Operationen.....	36
3.3.5.3 Zuweisungen und Throw Expressions.....	36
3.3.6 Schwer durchschaubare automatische Verhaltensweisen.....	37
3.3.6.1 Special Member Functions.....	37
3.3.6.2 Overloading mit Vererbung, Shadowing.....	38
3.3.6.3 Zweierlei Enums.....	38
3.3.7 Schlechte & Schreibintensive Namensgebungen.....	39
3.3.7.1 Datentypen.....	39
3.3.7.2 C++ und C-Style Casts.....	39
3.3.7.3 Raw- & Smartpointer.....	40
3.3.7.4 Funktionale Programmierung.....	41
3.3.8 Keyword Wiederverwertung.....	43
3.3.9 East Const und West Const.....	43
4 Implementierung.....	44
4.1 Umsetzung.....	44
4.1.1 C++ predigen, aber C# und Java trinken.....	44
4.1.2 Alternative Umsetzungsmöglichkeiten.....	44
4.2 Details.....	45
4.3 Lexer / lexikalischer Scanner.....	45
4.4 Parser / syntaktische Analyse.....	47
4.5 Sema / Semantische Analyse.....	49
4.6 Details zur Implementierung.....	52
4.6.1 Deklarationen.....	52
4.6.2 Statements.....	53
4.6.3 Expressions.....	54
4.6.4 Was lief nicht so gut (Überschrift verbessern???).....	54

4.7 Generator.....	54
5 Auswertung & Zusammenfassung.....	56
5.1 Bewertung der Grundsatzehaltung.....	56
5.1.1 Erwarte keine Wiederholung vom Benutzer.....	56
5.1.2 Außergewöhnliches Verhalten muss Explizit sein.....	56
5.1.3 Das was man häufig will kann Implizit sein.....	56
5.1.4 Breche nicht mit der Semantik von C++.....	56
5.1.5 Breche mit C sofern es einen Nutzen bringt.....	56
5.1.6 Entwickle die Syntax so weiter das sie Eindeutig bleibt.....	56
5.1.7 Sei auch einmalig Nützlich.....	57
5.1.8 Spare nicht mit neuen Keywords.....	57
5.2 Das sieht ja gar nicht mehr wie C++ aus!.....	57
5.3 Was konnte nicht behandelt werden.....	58
5.3.1 Alles was der C++ Semantik Widerspricht.....	58
5.3.2 Adresse-Von- und Indirektions-Operatoren.....	58
5.4 Performance Benchmark.....	59
6 Ausblick.....	61
6.1 Benamte Parameter.....	61
6.2 Intelligentere Parameter Übername.....	61
6.3 Ranges.....	62
6.4 Rest.....	62
7 Appendix.....	64
7.1 Literaturverzeichnis.....	64
7.2 Tabellenverzeichnis.....	65
7.3 Abbildungsverzeichnis.....	65
7.4 Listings.....	65
7.4.1 Vokabular des Lexers.....	65
7.4.2 Grammatik des Parsers.....	68

1 Einleitung

1.1 Definitionen

1.1.1 Formatierung dieses Dokumentes

Normaler Text mit einem *"Eingebetteten Zitat"* und dann weiter mit etwas höchst **Wichtigem**.

*"Dies ist ein Blockzitat
Von einer im Text genannten Person"*

Exemplarisch eine Codedatei mit dem Namen *myfile.cpp*:

```
// Dies hier ist Code  
int a;  
class B {...};
```

In Beispielcode werden Keywords **fett** hervorgehoben und eingebaute Typen *kursiv*. Hier wird auf die Variablen *a* und *b* Bezug genommen und die Typen *int* und *float* erwähnt, stets im Singular. Wenn Code im Fließtext geschrieben wird dann sieht es so aus `int a[] = {1,2,3};`.

1.1.2 Begriffsdefinitionen

In jedem Dokument, welches über mehrere Programmiersprachen schreibt, müssen zunächst Mehrdeutigkeiten und Kollisionen ausgeräumt werden. Besonders Erwähnenswert sind Unterschiede zur Namensgebung in C++.

In dieser Arbeit werden oft die englischen Varianten von Begriffen aus der Informatik / Programmierung genutzt, ein Beispiel zur Verdeutlichung: Ein **Feld** ist etwas wo der Bauer sein Korn anbaut, wo ein Mannschaftsspiel stattfindet oder eine Datenstruktur welche mehrere gleichartige Daten beherbergt. Im Kontext der Objektorientierung, in Datenbanken und Formularen werden jedoch auch Skalare Daten / Variablen *Feld* (engl. *Field*) benannt. Ein Ausdruck mit wesentlich mehr Trennschärfe, welcher das drittgenannte bezeichnet, ist **Array**. Das Wort **Ausdruck** aus dem letzten Satz hat eine Bedeutung, Ausdruck ist auch dass was aus einem Drucker kommt, eine **Expression** hingegen heißt zwar *Ausdruck* in Englisch, macht aber klarer dass es sich um einen Ausdruck im Kontext einer Programmiersprache handelt. Aus Gründen der Trennschärfe und weil sie im Alltag auch so verwendet werden, wird in diesem Dokument eben die englische Variante genutzt, aus Gründen der Lesbarkeit und Homogenität mit deutscher Großschreibung.

Namen von Keywords und Sprach-Konstrukten werden nicht pluralisiert.

C++

Wenn über C++ geschrieben wird, dann ist oft beides gemeint: die *Sprache C++* an sich und die *C++ Standard Library*. Diese sind, insbesondere seit C++11, sehr eng miteinander verbunden; so wurde die Sprache erweitert um bestimmte Funktionalität in der Bibliothek zu ermöglichen. Ebenso sind Komponenten der Bibliothek Notwendig um die Sprache vollständig zu nutzen.

Beispiele: `std::shared_ptr`, `std::initializer_list`, `std::move`, `std::dynamic_cast`

OOP

Objektorientierte Programmierung

Konstrukt

Alles innerhalb einer Programmiersprache: Deklarationen, Definitionen, Statements, Expressions, Attribute, Typen, Variablen. Erklärung dieser Begriffe folgend.

Keyword - Schlüsselwort

Wort mit spezieller Bedeutung innerhalb der Sprache welches nicht für eigene Namen verwendet werden kann.

Funktion & Prozedur

Manche Programmiersprachen unterscheiden hart zwischen diesen zwei Unterprogramm-Typen. (Als *Unterprogramm* versteht sich hier nicht die Definition welche COBOL verwendet)

Zum Beispiel liefern in Pascal *Prozeduren* keine Rückgabewerte. Außerdem wird in vielen Sprachen in *Funktionen* das Resultat nur aus den Eingaben erzeugt und hat keine Nebeneffekte, dieser Spezialfall wird hier "*pure Funktion*" genannt. In C++ werden all diese Unterprogramme *Funktionen* genannt, da es dort aber kein Keyword dafür gibt, konnten Benutzer im Sprachgebrauch natürlich ihre eigene Namensgebung verwenden.

In diesem Dokument wird sich der Gepflogenheit von C++ bedient: Alles sind Funktionen und können, soweit nicht anders annotiert, Nebeneffekte haben und auch keine Rückgabe haben.

{{{Myll führt Keywords für Funktionen ein und bietet beide Namensgebungen um beide Lager abholen. Diese können im fast synonym genutzt werden, analog zur Einführung von *class* zu *struct* in C++.}}}

Methode

Ist eine Funktion die mit einem Objekt verknüpft ist.

FhO - Funktion höherer Ordnung

Beschrieben im Kontext einer Prozeduralen Programmiersprache:

Eine FhO ist eine Funktion, welche eine weitere Funktion (oder auch Lambda) und eine Kollektion als Parameter übergeben bekommt. Die übergebene Funktion wird auf jedes Element der Kollektion angewendet. Diese Kollektion kann auch durch einen Iterator repräsentiert werden.

Map & Reduce

Map ist eine FhO, bei der pro Ursprungselement ein Zielelement erzeugt wird, welche wiederum in einer Kollektion gespeichert werden. In C++ ist dies durch `std::transform` implementiert.

Sie wird oft zusammen mit der FhO namens *Reduce* genutzt, welche aus vielen Ursprungselementen ein einzelnes Zielelement erzeugt.

GC - Garbage Collection

???

Template / templatisiert

Schablone / in eine Schablone verwandeln. Art der generischen Programmierung, Programmierung mit offen gehaltenem Typ.

TMP - Template Metaprogramming

Alternative Art der Programmierung bei der die Logik zur Kompilationszeit ausgewertet wird.

UB – Undefined Behaviour

Undefiniertes Verhalten. Tritt auf wenn der Programmierer die Regeln der Sprache verletzt. Nach dem Auftritt dieser Verletzung ist es dem Programm erlaubt jedweddes Verhalten zu zeigen. Bei dieser Art der Fehlfunktion muss ein C++ Compiler weder die Kompilation abbrechen, noch eine Warnung ausgeben.

Deklaration & Definition

Deklaration ist die Bekanntmachung eines Konstrukts, dessen Signatur.

Definition ist Implementation eines Konstrukts, dessen Innenleben.

RTTI

Runtime Type Information, ist die Bereitstellung von Typinformationen zur Laufzeit

DRY

Don't repeat yourself. ???

Syntactic Sugar

Einfachere Schreibweise für die es schon eine kompliziertere gibt.

Beispiel: `int_ptr[5]` ist das gleiche wie `*(int_ptr+5)`

und: `shape_ptr->draw()` ist das gleiche wie `(*shape_ptr).draw()`

Debug (Build)

Entwicklungsmodus einer Anwendung, üblicherweise nicht hoch-performant, dafür mit mehr Fehlerbehandlungsmöglichkeit erzeugt.

Release (Build)

Veröffentlichungsmodus einer Anwendung, üblicherweise mit allen verfügbaren Optimierungsmöglichkeiten erzeugt.

Expression

Ausdruck. In diesem Dokument im Englischen verwendet, weil es im Programmiersprach-Kontext viel eindeutiger ist als das Deutsche Wort. Ein Ausdruck ist für mich etwas was ein Drucker erzeugt.

Statement

Anweisung / Aussage. In diesem Dokument im Englischen verwendet, weil es im Programmiersprach-Kontext viel eindeutiger ist als das Deutsche Wort.

Declaration

Deklaration / Deklarierung. Inkonsequenterweise in diesem Dokument auf Deutsch genutzt.

Exceptions

Ausnahmen. Ein Mittel zur Fehlerpropagation innerhalb von Programmen, sehr oft in Objektorientierten Sprachen eingesetzt.

Member (Variablen/Funktionen)

Objektgebundene Variablen / Funktionen (Methoden)

(Operator) Precedence

Vorrangs Reihenfolge (von Operatoren).

Einfaches Beispiel aus der Mathematik: Punktrechnung vor Strichrechnung.

ABI – Application Binary Interface

Binärstandard. ???

RAII – Resource Acquisition Is Initialization

Ressourcenbelegung ist Initialisierung. ???

Polyfill

Füllfunktion. Welche ein Defizit in der Kern-Sprache oder der Standardbibliothek ausgleicht. Gängige Praxis bei JavaScript Frameworks, da jeder Browser andere Defizite hat.

Cacheline

???

1.2 Idee / Motivation

Die Inspiration zur Entwicklung dieser Sprache kam mir ursprünglich durch ein schon etwas älteres Dokument namens *'A Modest Proposal: C++ Resyntaxed'* von Werther und Conway. In Ihrem Dokument bemängeln sie zurecht einige Unschönheiten in der Grammatik von C++, welche zum Teil der Kompatibilität zu C geschuldet ist. Ihre syntaktische Reimaginierung von C++ nennen sie **SPECS**.

Auch der Erfinder von C++, Bjarne Stroustrup, glaubt

"within C++ there is a much smaller and cleaner language struggling to get out"

Sinngemäß ins Deutsche: *"in C++ schlummert eine wesentlich kompaktere und sauberere Sprache, welche Mühe hat ans Tageslicht zu kommen"*

was er in *'The Design and Evolution of C++'* schrieb, welches auch SPECS zitierte.

Ich teile diese Meinung. Wenn ich etwas in C++ umsetze, dann ist das Konzept in meinem Kopf meist wesentlich kompakter, als das was ich später umsetze.

Zwar hat modernes C++ (11/14/17) frischen Wind in die Sprache gebracht, aber trotzdem fast keine Zöpfe abgeschnitten. Auch wurden neue Sprachkonstrukte eingeführt, welche teils syntaktisch unschön und schreibintensiv sind, dem geschuldet, dass alter Code durch diese Ergänzungen nicht aufhören darf zu funktionieren.

Retrospektiv betrachtet hat die in SPECS vorgeschlagene Syntax nicht mehr viel mit C++ zu tun und dessen stark von Pascal angehauchte Syntax liegt fern von meinem Ästhetikempfinden. Mein Lebensweg als Programmierer ist von C und C++ ähnlichen Sprachen geprägt, weshalb die hier vorgestellte Lösung weit von der von SPECS abweicht.

Die zwei Hauptaufgaben von Myll sind nachfolgend beschrieben.

1.2.1 Alteingesessene C++ Programmierer abholen

Alteingesessene C++ Programmierer abholen und dessen Arbeit in Zukunft leichter zu gestalten.

Dies geschieht durch leicht zu merkende Sprachkonstrukte, wenig Wiederholung und Boilerplate???, enge Anlehnung an modernes C++ und Zukunfts-Offenheit für neue Funktionalität. Die Offenheit für zukünftige Sprachfeatures kann zunächst durch Polyfills erfolgen, sollte C++ jedoch später eine ähnliche Funktionalität anbieten, wird es einfach durchgereicht.

Auch der Ausstieg aus der Nutzung von Myll oder gar der einmalige Einsatz dessen ist möglich. Der erzeugte C++ Code soll leserlich und direkt weiter nutzbar sein. Kein geschriebener Code ist je vergebens.

1.2.2 Neulingen den Einstieg erleichtern

Neulingen den Einstieg in eine immer komplizierter werdende Sprache zu erleichtern.

Weniger Abweichung von anderen modernen Sprachen. Vereinfachung der Positionen an denen Keywords und Attribute auftauchen. Weniger überraschende und unbemerkte Nebeneffekte. Mehr selbstauferlegte Regeln mit denen man die Korrektheit seiner Implementierung überprüfen kann.

Durch die Erzeugung von lesbarem C++ Code kann dieser Transpiler auch beim lernen von C++ helfen, da man sich den erzeugten C++ immer anschauen kann.

1.3 Ziel

Diese zwei Hauptaufgaben stehen sich natürlich auch, teils extrem, entgegen. Es wird versucht einen gangbaren Mittelweg zu finden und Entschieden wird im Zweifel meist für die zukünftige Wartbarkeit und gegen die Vergangenheit.

Ein Ziel soll es auch die Betrachtung dessen sein, die Umsetzbarkeit einer Programmiersprache mit ähnlicher Komplexität wie C++ zu handhaben. Welche Aspekte von C++ sind problematisch und lösbar? Sind ähnlich positive Aspekte zu erzielen wie es Transpilation bei Webentwicklungen hervorgebracht hat?

Erklärung

Unterstützung bei der Einhaltung von Code-Richtlinien: Es gibt viele Richtlinien an die sich ein Team halten kann, aber diese Regeln nicht unabsichtlich zu verletzen ist nicht so einfach. Hier kann Myll einem unter die Arme greifen und diese Regeln für einen Überprüfen. Ein paar dieser Regeln sind vordefiniert, es soll in Zukunft auch möglich sein eigene Regeln zu definieren.

Lösung

Das Entscheidungsproblem unter dem C++ leidet, ob etwas eine Variablendefinition, eine Funktionsdeklaration oder eine mathematische Expression ist lässt sich durch die Einführung der *var* und *func* Keywords lösen. Es macht es auch fürs Menschenauge leichter entscheidbar was man gerade liest. Siehe 3.3.2.1

In Voraussicht wurde der Right-Shift-Operator in der Grammatik als zwei separate `>` definiert, welches die Probleme beim schließen zweier Templates löste. Ein Problem mit Templateargumenten trat dennoch auf, das etwa diese Expression hier falsch *verstanden* wurde:

```
static_cast< Type >( lhs ) | static_cast< Type >( rhs );  
static_cast<(Type >( lhs ) | static_cast< Type)>( rhs ); // eine große Expression
```

Der Ursprung war aber ein komplett anderer als die alte Template Problematik, hier war nämlich definiert das ein Templateargument eine beliebige Expression sein konnte und diese Verhält sich *greedy* solange noch ein schließendes `>` vor dem nächsten `;` auftaucht.

Die Problematik eine so komplexe Sprache wie C++ zu lesen, zu verstehen und dann wieder auszugeben ist ein hartes aber machbares Unterfangen. Ohne mir schon vor dieser Arbeit viel über C++ und allgemeinem Sprachdesign angeeignet zu haben, wäre es aber nicht in dem Zeitrahmen dieser Arbeit machbar gewesen.

Jemand anderes der ausgehend von dieser Arbeit einen Transpiler nach C++ aber auch C#, Java, Rust oder ähnlichem schreiben will, wird keine großartigen Probleme haben, solange die Grundstruktur der Sprache übereinstimmt.

1.4 Grundlagen

1.4.1 Was macht C++ besonders?

Es gibt sehr viele neue und moderne Programmiersprachen die vielfältige Einsatzgebiete haben, einfacher und sicherer zu benutzen sind und in Teilbereichen performanter als C++ sind. Die Popularität von C++ war mal auf einem Absteigenden Trend, dieser setzt sich aktuell aber nicht mehr fort.

Einsatzgebiete in denen häufig C++ für die Umsetzung verwendet wird:

- Betriebssysteme

- Treiber
- Embedded Computing
- Computerspiele
- Simulationen
- Statistik

In diesen Bereichen wird C++ entweder aufgrund seiner guten Performance oder wegen der geringen Größe der Laufzeitumgebung und des Kompilats eingesetzt. C++ war früher auch als Entwicklungssprache für Desktopanwendungen verbreitet, wurde in diesen Bereichen aber in weiten Teilen von Java, C# aber auch von Webanwendungen abgelöst.

Wieso hält sich C++ in den genannten Bereichen so hartnäckig?

- Maximale Kontrolle über den Speicher
- Keine erzwungene Garbage Collection (GC)
- Keine / schlanke Laufzeitumgebung
- Native Binärdateien
- Nichts bezahlen, was nicht genutzt wird
(abschaltbar: RTTI, Exceptions, Bounds/Overflow checks, etc)
- Keine erzwungene virtuelle Vererbung
- Mehrere Jahrzehnte alter C und C++ Code läuft fast immer ohne Anpassung

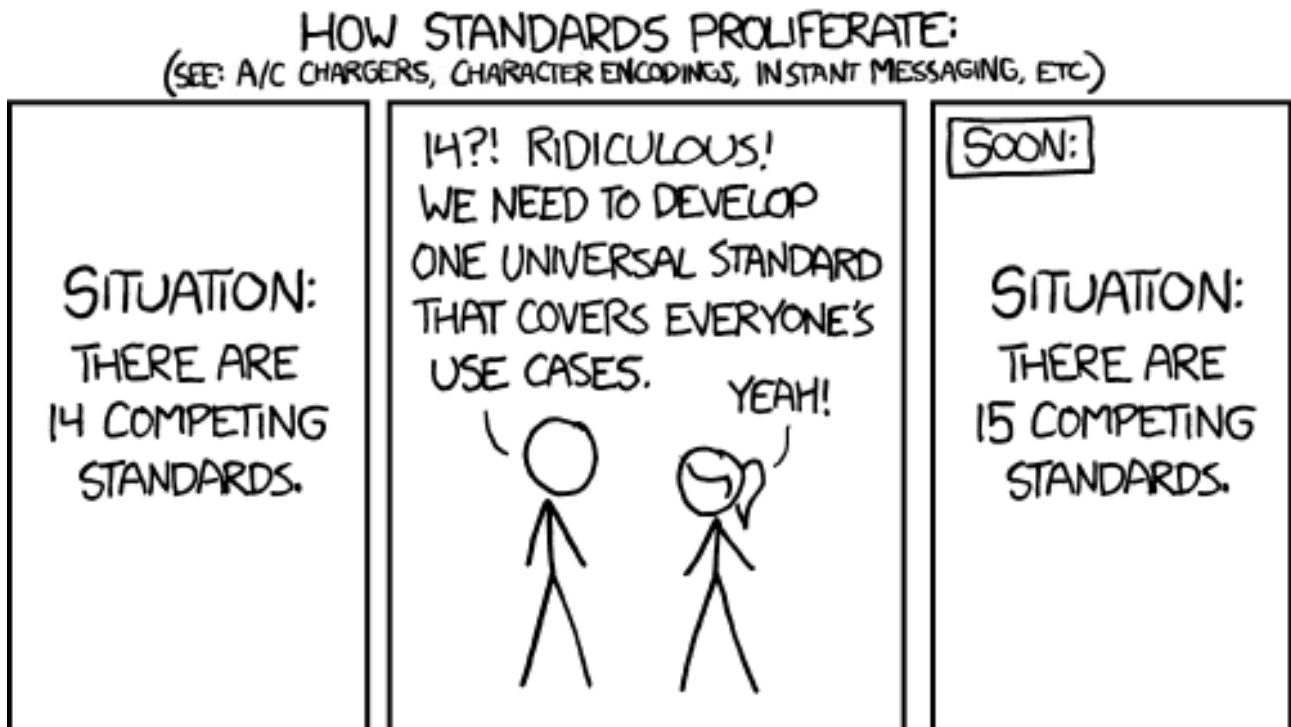
Mehr hierzu in 2.1.

1.4.2 Wieso noch eine neue Programmiersprache?

Warum eine neue Programmiersprache wenn C++ doch seine Bereiche so erfolgreich bedient?

Weil es sich mit vielen der anderen hier betrachteten Sprachen (siehe 2.1) einfacher entwickeln lässt als mit C++.

Trotzdem wären viele Sprachkonstrukte, sobald man sie auf Sprachebene und nicht nur auf Bibliotheksebene angehen kann, eigentlich leicht übertragbar auf C++.



<https://xkcd.com/927/>



2 Analyse

2.1 Die Besonderheiten von C++

Im vergangenen Kapitel wurde die besondere Stellung von C++ schon erwähnt, hier soll drauf eingegangen werden was dies in Detail bedeutet.

In C++ hat man maximale Kontrolle über den Speicher und kann auch selektiv, in transparenter Form, diese Kontrolle aufgeben, ohne auf Performance zu verzichten. Man kann einen Garbage Collector (GC) schreiben und diesen nutzen, man kann sich aber auch dagegen entscheiden und seinen Speicher manuell handhaben oder *Reference Counting* (Mitzählen wie oft ein Pointer verwendet wird) verwenden welches etwa vom `std::shared_ptr` bereitgestellt wird.

All diese Allokationen kann man entweder vom System Allokator durchführen lassen oder man kann einen spezifischen Allokator nutzen, zum Beispiel einen Arena-Allokator. Ein Arena-Allokator ist typischerweise für ein Subsystem zuständig. Dieser alloziert beispielsweise einen größeren Speicherblock vom Systemallokator, verwaltet und vergibt dann Stücke von diesem, teils gruppiert nach Größe (Binning), selbst. Dadurch sind beispielsweise gleich großen Objekte in einem zusammenhängenden Speicherbereich und im allgemeinen alle Objekte von ähnlichen Typen nah bei einander. Systemcalls werden dadurch seltener und interne Fragmentierung bekommt eine Obergrenze in Höhe der Bin-Größenabstände.

<https://www.youtube.com/watch?v=CVbnjydW5M0>

Frame-Allokatoren können in Programmen genutzt werden welche in deterministischen Zeitabständen einen *Frame* (hier *Schritt, Zeitschritt pro Bildaufbau*) abgeschlossen haben. Dies können zum Beispiel Simulationen, Graphische Anwendungen oder Computerspiele sein. Ein Frame-Allokator funktioniert so dass ein relativ großer Speicherbereich einmalig entweder statisch oder auf dem Heap alloziert wird. Fordert jemand Speicher von diesem Allokator an, wird dieser einfach vom Beginn des noch verfügbaren Bereiches abgegeben. Ein explizites Freigeben des Speichers gibt es nicht, denn dessen gesamter Speicher wird am Ende des Frames freigegeben, indem dessen Zeiger wieder auf den Anfang des Speichers gesetzt wird. Frame-Allokiertes-Speicher hat natürlich eine wichtige Einschränkung, er darf nicht über Frame-Grenzen gehalten werden. Dieser Allokator hat quasi gar keinen Verwaltungs-Overhead.

In einer anderen Sprache, welche auf das Vorhandensein einer GC aufbaut, ist all dies nicht mehr, oder nur partiell und mit hohem Aufwand möglich.

Strukturen besitzen in C++ ein fest vorgegebenes Speicherlayout, welches einem ermöglicht genau zu wissen ob sie etwa in eine *Cacheline* passen. Dadurch lässt sich sehr deterministisch die Laufzeit von darauf agierenden Algorithmen abschätzen.

Die aus C++ Code kompilierten Programme sind native Binärdateien, welche keine bzw. schlanke Laufzeitumgebungen mit sich bringen. Dies macht sie sehr schnell im Startup.

In C++ muss nichts bezahlt werden was nicht genutzt wird, dies beinhaltet sogar Kernfunktionen wie RTTI, Exceptions, Bounds-/Overflow checks, etc. Auch kann man sich aussuchen ob man virtuelle Methoden und Vererbung verwenden will oder ob man ohne auskommt, besonders bei kleinen Strukturen kann der Overhead enorm sein, siehe Kasten.

```
struct C { char a; };
struct VC { char a;
           virtual ~VC(){} };
sizeof( C ) == 1;
sizeof( VC ) == 16;
```

Mehrere Jahrzehnte alter C++ Code läuft fast immer, selbst die meisten alten, wie auch neuen, C Programme laufen ohne Anpassung.

Die Infrastruktur in Form der verfügbaren Compiler greift auf die Reife von jahrzehntelanger Optimierung zurück. Und dies ist nicht nur der Fall bei kommerziellen Produkten, denn es gibt mehrere kostenlose Compiler, viele davon sogar Open-Source: G++ der GNU Compiler Collection, Clang++ das C++ Frontend von LLVM und auch, durch Lizenzbedingungen eingeschränkt, der Microsoft Visual C++ Compiler.

2.2 Vergleich von C++ mit anderen Sprachen

Als Kompilationsziel von Myll soll zwar C++ dienen, jedoch mag sich Myll zum einen Syntaktisch und zum anderen von den Features anderer Programmiersprachen inspirieren lassen. Aus diesem Grund sei hier C++ mit einigen ausgewählten Programmiersprachen verglichen.

2.2.1 Vergleich mit C

Der Vorgänger von C++.

- + Viel Kontrolle, man sieht genau was passiert
- + Viel simpler als C++
- + Sehr viele Freiheiten, aber auch Verantwortung bei der Speicherverwaltung
- Schlechte Datenkapselung / Keine native Unterstützung für OOP
- Unübersichtlich bei großen Projekten, denn ohne Namespaces landet alles im globalen Raum
- Es ist der Ursprung einiger in dieser Arbeit betrachteten syntaktischen Schwächen von C++
- Mehr LOC (Lines of Code, Codezeilen) für die Lösung des gleichen Problems

Hier ist ein Beispiel welches Aufzeigt was ich mit *viel Kontrolle* und *viel simpler* meine. Es ist Code welcher so in C, als auch in C++ auftreten kann. Die Kommentare beschreiben was in C++ zusätzlich, zum offensichtlichen passieren kann.

```
{  
    Foo f;           // kann in C++ einen potentiell teuren Konstruktor von Foo aufrufen  
    A * a = f.get(); // kann in C++ eine Exception werfen und den folgenden Code überspringen  
    a->update();     // kann in C++ eine virtuelle Methode aufrufen  
}
```

<https://flooooh.github.io/2018/06/02/one-year-of-c.html#less-language-feature-anxiety>

2.2.2 Vergleich mit C# und Java

C# und Java sind als designierte Nachfolger von C++ mit einer Transfusion dessen OOP Bluts entstanden. Sie haben sehr viel untereinander gemein und werden deshalb hier gemeinsam betrachtet. ??? Recherche

- + Es ist leichter fehlerfreien C# und Java Code zu schreiben
- + Fehlermeldungen in Generics/Templates beschreiben viel präziser den Fehlerfall
- + Einfach auf vielen Architekturen ausführbar (oft ohne Neukompilierung)
- + Weniger LOC für die Lösung des gleichen Problems
- + Viele äquivalente Features viel früher verfügbar als in C++ (concepts, auto, ranges, override, ...)
- Erzwungene GC, weniger Kontrolle über den Speicher
- Keine multiple Vererbung
- Nicht abschaltbare Bounds/Overflow checks, Reflection, Exceptions
- Ein leichtes Performance Defizit welches bei Desktopanwendungen verschmerzbar ist

2.2.3 Vergleich mit D

D hatte sich vorgenommen die Reihe der Sprachen BCPL \rightarrow B \rightarrow C \rightarrow C++ \rightarrow D fortzusetzen.

Leider konnte D nicht annähernd den Erfolg von C oder C++ erreichen, obwohl vieles von C++ gelernt und womöglich besser gemacht wurde.

- + Sehr nah an C++ angelehnt
- + Weniger LOC für die Lösung des gleichen Problems
- GC für die Nutzung der Standard Bibliothek nötig, weniger Kontrolle über den Speicher
- Keine multiple Vererbung

2.2.4 Vergleich mit Rust

Eine sehr neue Programmiersprache welche ihren Fokus auf die Sicherheit dessen was maschinell überprüfbar ist legt. Die statischen Überprüfungen welche der Rust Compiler vornimmt sind in der Ausprägung noch nicht verbreitet gewesen.

- + Stärkere statische Checks welche zu wesentlich sicherem Code führen sollen
- + In Zukunft ist Rust ein Contender um einige Bereiche welche C++ heute bedient [TR-R5Y]
- (Anfänglich) schwerer zu schreiben durch Paradigmenwechsel des Lifetime-Managements
- Exorbitante Kompilationszeiten ???

2.2.5 Vergleich mit Ruby

Dieser Vergleich ist ein wenig weiter her geholt als die anderen betrachteten Sprachen. Ich wollte hier dennoch gern über den nahen Tellerrand hinausblicken um mich auch von weiter weg inspirieren zu lassen.

- + Unglaublich wenig Code ist nötig um simple Probleme zu lösen
- + Lambdas sind schon sehr lange ein gut integrierter und fast unsichtbarer Teil der Sprache
- Interpretierte Sprache (langsam in der Ausführung)
- Erzwungene GC, weniger Kontrolle über den Speicher
- Keine Typsicherheit

2.2.6 Vergleich von JavaScript mit TypeScript

Dieser Vergleich findet nicht mit C++ statt, sondern mit der Sprache auf der TypeScript basiert. Er dient als das Beispiel eines Dreisatzes:

Was TypeScript für JavaScript ist, soll Myll für C++ sein!

Das ist *noch* kein Zitat, weil ich es selbst gesagt habe. Ich finde es nur sehr einfach auf den Punkt gebracht was das Ziel dieser Sprache sein soll.

JavaScript leidet unter dem gleichen Problem unter dem auch C++ leidet, nämlich das alter Code weiterhin so funktionieren muss wie bisher. Demnach können sich beide Sprachen nur in den Dimensionen ausdehnen, welche zuvor kein kompilierendes/lauffähiges Programm ergeben hätte. Dadurch sind sie verdammt ihre schlechten Entscheidungen für immer mit sich herumzutragen.

Hier zunächst der gut lesbare **TypeScript Code** welchen wir folgend in JavaScript übersetzen.

```
class Human {
  _name: string;
  constructor(name: string) {
    _name = name;
  }
}
class Student extends Human {
  _mtknr: number;
  constructor(name:string, mtknr:number) {
    super(name);
    _mtknr = mtknr;
  }
  getGrade() : string {
    return "A+";
  }
}
```

Daraus macht der TypeScript Transpiler diesen ES5 **JavaScript Code**:

```
var __extends = /* gekürzt: 12 Zeilen lange Polyfill Funktion */
var Human = /** @class */ (function () {
  function Human(name) {
    this.name = name;
  }
  return Human;
})();
var Student = /** @class */ (function (_super) {
  __extends(Student, _super);
  function Student(name, matrikelnummer) {
    var _this = _super.call(this, name) || this;
    _this.matrikelnummer = matrikelnummer;
    return _this;
  }
  Student.prototype.getGrade = function () {
    return "A+";
  };
  return Student;
}(Human));
```

Man müsste ähnlichen Code schreiben, wenn man will das er auf einem ES5 tauglichen Browser läuft.

TypeScript bietet viele Vorteile gegenüber purem JavaScript.

- Es ist leichter Les- und Schreibbar
- Es bietet optional Typsicherheit
- Es ermöglicht Funktionalität von moderneren JavaScript Versionen, in älteren Versionen zu nutzen
- Es erlaubt direkt bei der Übersetzung eine Aggregation von mehreren Dateien in eine

2.3 Problematische Aspekte von C++

Einige der problematischen Konstrukte in C++ sind der direkten Kompatibilität mit C geschuldet, andere sind von C++ so eingeführt worden. Die Betrachtung dieser Probleme findet sich Übersicht halber direkt gepaart mit deren Lösung in Kapitel 3.3 auf Seite 21.

3 Konzept

3.1 Einleitung

Um bei der Konzeption der Sprache das Ziel nicht aus den Augen zu verlieren, wurden einige simple Grundsätze definiert, welche auch auf Englisch verfügbar sind. Die Einhaltung der anfangs aufgestellten Grundsätze soll später bewertet werden.

Der Autor vertritt die Auffassung das viel Inspiration von anderen Sprachen eine gute Sache ist und mehr Wiedererkennungswert den Einstieg einfacher macht.

3.1.1 Grundsätze der Sprache Myll

1. Erwarte keine Wiederholung vom Benutzer
2. Außergewöhnliches Verhalten muss Explizit sein
3. Das was man Naïv erwartet und keine Nebeneffekte hat kann Implizit sein
4. Breche nicht mit der Semantik von C++
5. Breche mit C sofern es einen Nutzen bringt
6. Entwickle die Syntax so weiter das sie Eindeutig bleibt
7. Sei auch dann nützlich, wenn die Entwicklung am einmalig erzeugten C++ Code weitergeht
8. Spare nicht an neuen Keywords, wenn die Lesbarkeit profitieren kann

3.1.2 Fundamental Principles of Myll

1. Don't ask the user to repeat themselves
2. Exceptional behavior needs to be explicit
3. What you naïvely expect to happen and has no side-effects can be implicit
4. Don't break with C++'s general semantic
5. Do break with C if there is a benefit
6. Evolve the syntax that it's unambiguous
7. Be useful even if a one-time translation to C++ is all that's wanted
8. Don't be frugal with new keywords, if it benefits readability

3.1.3 Inspiration

Ich habe mich schon eine lange Zeit vor dem Beginn dieser Arbeit mit dem Design und der Erstellung von Programmiersprachen beschäftigt. Das Video »**Ideas about a new programming language for games**« von **Jonathan Blow**, aus dem Jahre 2014, brachte mich dazu mein Vorhaben, eine eigene Programmiersprache zu erstellen, in die Tat umzusetzen. Dieses und viele seiner weiteren Videos diente auch als Inspirationsquelle vieler Aspekte der hier vorgestellten Sprache. Jonathan Blow ist als Game-Designer und Chef-Entwickler der Spiele *Braid* und *The Witness* bekannt. Er entwickelt aktuell parallel die Programmiersprache **Jai**, zusammen mit einem weiteren Spiele-Projekt, einem komplexen Sokoban Puzzlespiels, welches in eben dieser Programmiersprache implementiert ist.

Auch ein zweites Projekt begleitete den Pfad, der zur Erstellung dieser Arbeit führte, über eine längere Zeit: »**Bitwise**« von **Per Vognsen**. In Bitwise ging es darum einen kompletten Computer, also den gesamten Hard- und Software-Stack, als Open-Source selbst zu erzeugen. Von besonderem Interesse für mich war die Erzeugung des **Ion** getauften Compilers dieses Projektes. Herausstechend daran war das die Entwicklung live übertragen wurde und der Code dieser Entwicklung jederzeit einsehbar war. Zu sehen, dass ein Compiler innerhalb kurzer Zeit von seiner Konzeption zur Funktion gebracht werden konnte war höchst ermutigend. Auch beim Sprachdesign wurde ich Inspiriert, eine einfach lesbare Syntax, sowohl für Parser, als auch für Menschen zu erzeugen. Per Vognsen war lange Zeit als Spielentwickler und Systemprogrammierer tätig, welcher aktuell seine Zeit in informative und lehrreiche Projekte mit geringem kommerziellem Interesse investiert.

Konträr zu diesen beiden Projekten liegt mein Fokus nicht auf extrem schneller Verarbeitung / Kompilation, sondern auf schneller Einarbeitung, weniger Flüchtigkeitsfehler und weniger Wiederholung.

Ein Artikel ist mir positiv und inspirierend aufgefallen »**Ideas for a Programming Language Part 3: No Shadow Worlds**« von **Malte Skarupke**, welcher mein Bewusstsein auf die unterschiedlichen Programmebenen hinwies, welche alle gleichzeitig in einer einzelnen Programmiersprache vorhanden sind. Da ist das erwartete Prozedurale als auch das Objektorientierte in C++. Dahinter verborgen, quasi im Schatten dieser Offensichtlichen, liegen noch Generisch (und dessen Spezialfall TMP), Funktional und vom Präprozessor vorverarbeitet. Das Schlimme daran ist das sich teils mehrere dieser schattenhaften Ebenen überlagern und dadurch das Verständnis dessen was an einem Codeblock passiert verkomplizieren. Auch aufgrund dieses Artikels habe ich mich strikt gegen die Aufnahme eines Präprozessors entschieden und versuche TMP so gut es geht durch einfachere Lösungen zu ersetzen.

Außerdem stieß ich bei der Recherche dieser Arbeit auf ein Proposal (offizielles Vorschlags Dokument) welches viele problematische Aspekte von C++ auflistet und für deren Behandlung den offiziellen Weg durch den Standardisierungsprozess gehen will. Es handelt sich dabei um »**Epochs: a backward-compatible language evolution mechanism**« von **Vittorio Romeo**. Das Proposal zeigt, wie die Probleme von C++ durch einen "Epochen-Schalter" gelöst werden können. Dieser Schalter könnte selektiv, pro Datei / Translation-Unit sogar "Zöpfe abschneiden". Sollte dieses oder ein ähnliches Proposal die Aufnahme in C++ finden, werden bestimmt nicht alle der betrachteten Probleme gelöst und der frühestmögliche Zeitpunkt wäre 2023 oder gar 2026. Leider sehe ich auch einige der Richtlinien unter 8.12 als absolut nicht *unumstritten* an und ich kenne einen ganzen Schlag Leute die dies ähnlich sehen (Spielentwickler).

Analysis – Paralysis???

3.1.4 Lerneffekte der verglichenen Sprachen

Welche Lerneffekte kann Myll aus den mit C++ verglichenen Programmiersprachen ziehen?

Von **C** kann man Simplizität Lernen: einfache Syntax für Casts, an Raw-Pointer angelehnte Smartpointer Syntax. Syntaktisch schwer durchschaubare und problematische Konstrukte hingegen sollen vermieden werden. Von **C#** und **Java** kann man sich einen Feature-Vorsprung zu C++ und eine entschlackte Syntax anschauen. Das Ziel von **D** war ein recht ähnliches wie meines, nämlich ein besseres C++ zu sein, Myll wird keine GC verwenden und auch nicht (absichtlich) auf Kernfeatures von C++ verzichten. Von **Rust** direkt schneidet sich Myll keine Scheibe ab, einen Paradigmenwechsel in der Schreibweise von C++ soll es explizit nicht geben, es wird lediglich versucht den halben Weg in Richtung sicherer Software durch die Behandlung einfach zu lösender Probleme zu gehen. Von **Ruby** soll sich ein "Das könnte doch viel einfacher geschrieben werden!" abgeschaut werden, welchen C# sich schon über die Jahre von Ruby oder ähnlichen Sprachen abgeschaut hat. Von **TypeScript** kommt die ursprüngliche Umsetzungsidee dieses Transpilers.

3.2 Design der Sprache Myll

Im nächsten Abschnitt werden die angesprochenen Probleme in C++ behandelt, aber eine schlichte Behandlung einzelner Probleme macht noch keine Sprache aus, sie bietet lediglich ein besseres Fundament. Eine Programmiersprache sollte auch ein einheitliches Gesamtbild, wenig widersprüchliche Konstrukte und eine gute Lesbarkeit aufweisen.

3.2.1 Neue Keywords

Myll nutzt einige neue Keywords, denn Syntax ohne Keywords hat sehr schnell alle *schönen* Konstellationen verbraucht und Monster wie diese hier entstehen:

```
[](){}();           // valides C++11, leeres Lambda welches direkt aufgerufen wird
void(*ffp)(int*)(int,int)); // valide seit dem Anbeginn von C++
```

Die Liste der neuen Keywords ist:

```
func, proc, method, ctor, dtor // alle Arten von Funktionen
var, field                     // alle Arten von Variablen
get, set, refget               // zur Erstellung von Accessoren
import, module                 // Bekanntmachung und Inkludieren von Modulen
loop, times                    // neue Schleifen
requires, alias                 // etc
```

Es wurden Abkürzungen mittlerer Länge gewählt welches sich an die ursprünglichen Keywords aus C++ anlehnt: struct(ure), enum(eration), float(ing point number), double (precision float), int(eger). Für Funktionen wäre *fn* zu wenig Information und kollidiert zu leicht.

Außerdem konnten viele Keywords aus C++ im allgemeinen Programm freigegeben werden, diese haben in Myll nur noch als Auszeichnung innerhalb von Attributen eine Bedeutung:

```
static, include, export, virtual, override, final, ...
```

Da Myll jedoch im aktuellen Stand nach C++ übersetzt, sollten diese natürlich noch nicht für Identifier verwendet werden.

3.2.2 Übersicht der Syntax von Myll

Jede Deklaration beginnt mit einem Keyword welche es direkt erkennbar macht was die Zeile enthält. Was darauf folgt ist meist ein Name oder auch ein Typ, gefolgt von weiteren Informationen welche entweder mit Semikolon oder einem Block (Geschweifte Klammern und deren Inhalt) beendet werden. Dies gilt hier auch für *import* und *module*, wird in ??? erklärt.

Die Syntax der Deklarationen von Variablen und Funktionen sei Vorab kurz erklärt um ein Grundverständnis aufzubauen. Variablen werden mit dem neuen Keyword *var*, *const* oder *field* eingeleitet und verhalten sich danach weitestgehend wie C++, mehr dazu in 3.3.2.1. Funktionen werden mit dem neuen Keyword *func*, *proc* oder *method* eingeleitet, die Rückgabetypen werden Folgend angegeben, mehr dazu in 3.3.4.1.

Der die übrigen Aspekte der Syntax von Myll werden im Laufe des Kapitels 3.3 erklärt.

3.2.3 Gesamtbild

Dieses Beispiel braucht noch nicht vollständig verstanden zu werden, es ist nur da um ein Grundeindruck zu vermitteln. Trotzdem etwas Erklärung zu dem was zu sehen ist: Jede Deklaration beginnt mit einem Keyword welche es direkt erkennbar macht was die Zeile enthält. Was darauf folgt ist meist ein Name oder auch ein Typ, gefolgt von weiteren Informationen welche entweder mit Semikolon oder einem Block (Geschweifte Klammern und deren Inhalt) beendet werden. Dies gilt hier auch für *import* und *module*, später erklärt in ???.

```

module MyContainers;

import IOStream;

func main( int argc, char*[] argv ) -> int {
    using std;
    var MyStack<i8> stack;
    stack.push( 42 );
    if( !stack.empty() ) {
        cout << stack.top() << endl;
        stack.pop();
    }
    return 0;
}

class MyStack<T> {
    const usize _default_reserved = 8;
    field usize _reserved;
    field usize _size = 0;
    field T[]*! _data;

    public:
    ctor( usize reserved = _default_reserved ) {
        _reserved = reserved;
        _data = new T[_reserved];
    }

    [const]
    method empty() -> bool
        ==> _size == 0;

    [const, precond(!empty())]
    method top() -> T
        ==> _data[_size-1];

    method push( T val ) {
        if( _size >= _reserved )
            grow();
        _data[_size] = val;
        ++_size;
    }

    [precond(!empty())]
    method pop() {
        --_size;
    }

    private:
    method grow() {
        const usize new_reserved = _reserved * 2;
        var T[]*! new_data = new T[new_reserved];
        do _size times i {
            new_data[i] = _data[i];
        }
        _reserved = new_reserved;
        _data = new_data;
    }
}

```

Hinweis zum Beispiel:

Das Member `_data` ist ein Unique-Pointer (*!), deshalb braucht diese Implementierung weder ein `delete` in `grow()`, noch einen manuell implementierten Destructor.

Auf interne Fehlerbehandlung mittels Exceptions wurde explizit verzichtet, da der Nutzer von außerhalb die üblichen Fehlerfälle vorab prüfen kann.

??? `precond` / `postcond` == `require` / `ensure`

3.3 Lösung der problematischen Aspekte von C++

Einige der problematischen Konstrukte in C++ sind der direkten Kompatibilität mit C geschuldet, andere sind von C++ so eingeführt worden.

Die Lösung der Probleme findet sich jeweils direkt im gleichen Absatz.

3.3.1 Deklarationen, Definitionen und Initialisierung

C++ verhält sich selbst in sehr simplen Beispielen wie eine *Single-Pass parsende* Programmiersprache (was sie nicht ist) und erfordert dadurch vom Programmierer ein hohes Maß an Aufmerksamkeit bei der Reihenfolge der Programmierung oder fordert die Wiederholung von bereits Niedergeschriebenem.

3.3.1.1 Reihenfolge und Prototypen

Dieser Code kompiliert so nicht, er wirft den Fehler das `a()` nicht deklariert wurde.

```
int main() {  
    a();  
}  
void a() {...}
```

Um diesen Fehler zu beseitigen müssen wir entweder den Code so umstellen das `a()` vor `main()` definiert wird oder wir fügen eine Prototypen-Deklaration für `a()` vor `main()` ein.

Im Falle eines alternierend rekursiven Funktionsaufrufs bleibt einem nur die Möglichkeit des Prototypen.

```
void partition(...);  
void qsort(...) { ... partition(...); ... }  
void partition(...) { ... qsort(...); ... }
```

Computer sind schnell genug diese Probleme für einen zu lösen. Lediglich der C als auch C++ Standard stehen der Unterstützung solch einfacher Inferenz im Wege.

Lösung

In Myll gibt es keine separate Deklaration und Definition, beides erfolgt in einem Schritt und die Reihenfolge ihrer ist Irrelevant. Sollte im erzeugten C++ Code dadurch ein Reihenfolge Problem entstehen, wird automatisch eine Prototyp-Deklaration vor der Nutzungsstelle erzeugt.

3.3.1.2 Aufteilung auf verschiedene Orte

Deklarationen, Definitionen und Initialisierung können entweder innerhalb einer Datei an verschiedenen Stellen und verschiedenen Schreibweisen vorkommen oder oft sogar über Dateigrenzen hinweg verteilt sein.

Dies macht den Code schwerer durchschaubar. Jegliche andere moderne Programmiersprachen erlaubt es einem (oder zwingen einen gar dazu) alles was z.B. zu einer Klasse gehört in einer einzelnen Datei abzulegen.

Dieses Beispiel soll einige der Probleme aufzeigen.

Datei *myclass.h*:

```
class MyClass {
    int myField1;
    int myField2;
    int myField3 = 3;           // Funktioniert seit C++11
    static int myStatic;       // Funktioniert nicht wie myField3
public:
    MyClass();
    int myMethod1();
    int myMethod2() {...}      // Kann inlined werden, implizit weil inline implementiert
    inline int myMethod3() const;
    virtual int myVirtual();
}
int MyClass::myMethod3() const {...} // Kann inlined werden, Keyword const wird wiederholt
```

Datei *myclass.cpp*:

```
#include <myclass.h>
int MyClass::myStatic = 1;    // Keine Wiederholung des Keywords static, Typ int wiederholt
MyClass::MyClass() : myField(1) {
    myField = 2;
}
int MyClass::myMethod1() {...} // Kann nicht ohne Weiteres inlined werden
int MyClass::myVirtual() {...} // Keine Wiederholung des Keywords virtual
```

In würde man in diesem Beispiel *myMethod1* mit `inline` auszeichnen, würde weder inline-fähiger Code entstehen, noch würde gewarnt das diese Auszeichnung keinen Nutzen hat.

Bewertung

- Manche Keywords müssen in Deklaration und Definition geschrieben werden, andere nicht
- Der Abschnitt "`MyClass::`" wird sehr oft wiederholt geschrieben
- Oft gibt es mehrere unterschiedliche Möglichkeiten das gleiche zu schreiben

Lösung

Durch diese Zusammenführung ist auch die Aufteilung auf separate Dateien unnötig geworden. Es werden natürlich für den erzeugten C++ Code weiterhin `.h` und `.cpp` Dateien erzeugt.

Das folgende Beispiel enthält die gleiche Menge an Information wie die Dateien *myclass.h* und *myclass.cpp* aus der Analyse. Es lassen sich aus diesem Beispiel die originalen Dateien (fast identisch) wiederherstellen. Das Attribut `inline` entscheidet hier ob die Implementierung in der `.cpp` oder der `.h` Datei landet, es hat immer eine Auswirkung.

Datei *myclass.myll*:

```
class MyClass {
    field int myField1;
    field int myField2;
    field int myField3 = 3;
    [static] field int myStatic = 1;
public:
    ctor() {
        myField1 = 1;           // Generiert wenn möglich C++ Code wie myField1 im Original
        myField2 = 2;           // Generiert wenn möglich C++ Code wie myField1 im Original
    }
    method myMethod1() -> int {...}
    [inline] method myMethod2() -> int {...}
    [inline,const] method myMethod3() -> int {...}
    [virtual] method myVirtual() -> int {...}
}
```

Analyse des Ursprungscodes und der bereinigten Syntax, ohne Whitespace und Kommentare:

Ursprung, C++: 18 Zeilen, 34 Wörter, 351 Zeichen

Bereinigung: 13 Zeilen, 26 Wörter, 264 Zeichen

Diese Transformation brachte in diesem sehr simplen, von Implementierung befreiten, Beispiel eine Reduktion der Zeilen um 28%, der Wörter um 24% und der Zeichen um 21%. Es geht in diesem betrachteten Punkt des Konzepts auch explizit nicht um die Implementierung. Größere und realistischere Beispiele werden späteren Verlauf der Thesis Betrachtet und Ausgewertet. ??? neu berechnen

3.3.1.3 Aufteilung im Kontext sich stetig verändernder Codebases

Die Verteilung auf mehrere Dateien ist besonders problematisch wenn ein C++ Projekt ständig signifikante Änderungen / Restrukturierungen erlebt. So muss im Beispiel das eine Methode `inline` fähig gemacht werden soll (selbst wenn die Logik unverändert bleibt) ihre komplette Implementierung aus der .cpp Datei ausgeschnitten und in die .h Datei eingefügt werden. Dadurch verliert man z.B. in seiner Versionsverwaltung den Bezug und kann nicht mehr einfach die Veränderungshistorie der Methode nachschlagen.

Das gleiche Problem tritt auf wenn eine Klasse generisch gemacht werden soll (mittels Templates), nur muss in diesem Fall die **komplette** Implementierung in die Header Datei wandern.

Lösung

Sich ständig wandelnder Code profitiert von der irrelevant gewordenen Reihenfolge von Deklarationen und der Vereinigung auf eine Datei. Ein Attribut hinzu oder ein Attribut entfernt, erzeugt eine geänderte Zeile im *diff* der Versionen, der C++ Code kann sich dadurch signifikant ändern.

Wenn in Myll eine Klasse generisch gemacht wird, fügt man einfach den Template-Parameter hinzu (und ändert natürlich die Vorkommen der zu ersetzenden Typen durch z.B. `T`).

3.3.2 Fehleranfällige Variablen & Typen Deklarationen

Wie findet die Deklaration von Variablen in C und C++ statt und welche Problematischen Situationen können dabei entstehen.

3.3.2.1 Ambiguität einer Deklaration

Sind die folgenden Statements Deklarationen oder nicht?

```
a * b;  
c d;  
c e();
```

Die Antwort darauf lautet: *Es kommt drauf an.*

Sollte *a* ein Typ sein, wird *b* als ein Pointer auf ein *a* Deklariert, ist *a* aber eine Variable wird `a.operator*(b)` aufgerufen. Der andere Fall ist zum einen *d*, welcher eine Variable vom Typ *c* deklariert und zum anderen *e* welcher grade nicht eine Variable vom Typ *d* deklariert die den Standardkonstruktor aufruft, sondern es ist eine Prototypen-Deklaration einer Funktion namens *e* welche ein *c* zurück gibt.

Diese Probleme bieten sich nicht nur dem Leser dieser Statements, sondern auch der Compiler musste aufwändiger erstellt werden, um diese Konstruktionen richtig interpretieren zu können.

Siehe: [LEXHACK]

Lösung

Um der Ambiguität ob etwas Beispielsweise eine Variable, eine Multiplikation oder einen Funktions-Prototyp ist entgegen zu wirken führt Myll Keywords zur einfachen Identifikation deren ein.

```

var a * b;      // Variablen Deklaration
const a * b;    // Konstanten Deklaration (mit fehlender Initialisierung)
a * b;          // Multiplikation
var c d;        // Variablen Deklaration
var c e();      // Variablen Deklaration mit Aufruf des Standard Konstruktors
func e() -> c;   // Funktionsdeklaration (unnötig in Myll, mehr dazu in dediziertem Kapitel)

```

Das Keyword *var* für Variablen, *const* für Konstanten (was in C++ bereits so aussieht), *func* für Funktionen (mehr zu Funktionen in 3.3.4.1).

Außerdem ist es gleichförmig zu anderen Konstrukten in C++, welche mit einem vorangestellten Keyword den Leser direkt wissen lassen was folgen wird: *class*, *struct*, *union*, *enum*, *namespace*, *using*, *static*; selbst *template*, *typename*, *typedef*, *#define* erfüllen dieses Kriterium.

Inspiriert war diese Syntax von einigen anderen Programmiersprachen, aber die Entscheidung diese Keywords zu verwenden kam durch das Projekt Bitwise von Per Vognsen, welcher in seiner Programmiersprache **Ion** die gleichen Keywords verwendet [ION].

3.3.2.2 Pointer, Referenz & Array

Pointer

Naiv hätte ich bei der Variablendeklaration von

```
int * ip, jp;
```

erwartet das *ip* und *jp* Pointer auf *int* sind. Damit lag ich (so wie viele meiner Kommilitonen und Kollegen) falsch und als ich zum ersten Mal über dieses Problem stolperte, stellte sich nach etwas Recherche raus das *ip* ein Pointer und *jp* nur ein *int* Objekt waren, richtig wäre an dieser Stelle

```
int *ip, *jp;
```

gewesen. Aus diesem Grund empfehlen viele Bücher das man nur eine Variable pro Zeile deklariert, aber grade jene Einsteiger, welche über dieses Problem Stolpern, lesen (wenn überhaupt) nicht zuerst *Best-Practice* Bücher.

Der Ursprung dieser Syntax ist in [KR-CPL] so beschrieben:

```

"The declaration of the pointer ip,
    int *ip;
is intended as a mnemonic; it says that the expression *ip is an int."

```

Das heißt dass man später durch Aufruf von **ip* an das bezeigte *int* gelangt. Kommt noch ein zweiter Variablenname hinzu, muss auch er mit der Eselsbrücke versehen werden wie man an sein *int* gelangt.

Obwohl ich diese Gedächtnisstütze jetzt schon lange kenne, arbeitet mein Hirn mit einem anderen mentalen Modell. In diesem Modell besteht eine Variablendeklaration aus einem Typ (der Indirektionsebenen beinhaltet) und einem oder mehrerer Namen. Nicht aus einem Basistyp (ohne Indirektionsebenen) und Namen vermischt mit Eselsbrücken wie ich an den Basistyp komme.

Als eines von vielen möglichen Beispielen von Template-Klassen verhalten sich die in C++11 eingeführten Smartpointer getreu meinem mentalen Modell, so das

```
unique_ptr<int> ip, jp;
```

jeweils *ip* und *jp* als (Unique) Pointer auf ein *int* definiert werden.

Referenz

Das gleiche Problem der Pointer gilt auch bei der Deklaration von Referenzen.

Sie haben nur den wichtigen Unterschied das die Mnemonik bei ihnen nicht mehr gilt, um aus einer mit

```
int &ir;
```

definierten Referenz das *int* abzurufen darf man nicht `&ir` schreiben, denn dadurch würde man einen Pointer auf das *int* erhalten. Dafür können Kernighan & Ritchie nichts, weil Referenzen erst in C++ Einzug hielten.

Arrays

Ein ähnliches Bild wie bei den Pointern findet sich bei der Syntax zur Definition von Arrays [KR-CPL]:

"The declaration

```
int a[10];
```

defines an array a of size 10, that is, a block of 10 consecutive objects named a[0], a[1], ... , a[9]."

Auch hier kommt man im Nachhinein durch Aufruf von `a[zahl]` an die jeweiligen *int*, interessanterweise ist aber genau der Aufruf von `a[10]` ein Fehler, welcher hinter das Ende des Array adressiert, welches UB hervorruft. Um zwei oder mehr Arrays zu erzeugen muss man dies hier schreiben: `int a[10], b[10];`

Das C++11 Äquivalent dazu ist `array<int,10> a, b;` bei dem wieder *a* und *b* jeweils Arrays der Größe 10 vom Typ *int* sind.

Lösung

In Myll ändert sich die Pointer, Referenz und Array Syntax so das der Stern, das Kaufmanns-Und sowie die eckigen Klammern zum Typ gehören und der Typ für alle Variablennamen gleich ist:

```
var int*    a, b; // Variablen a und b sind beide Pointer auf int
var int&    c, d; // Variablen c und d sind beide Referenzen auf int
var int[10] e, f; // Variablen e und f sind beide int Arrays mit 10 Elementen
```

Verloren geht dadurch die Möglichkeit total unterschiedliche Initialisierungen wie diese hier in einer Zeile schreiben:

```
int i = 1, *p = NULL, f(), (*pf)(double), a[10];
```

Diese Änderungen beseitigt nicht nur die potentiell Fehleranfällige Syntax von K&R C, sondern harmonisiert außerdem mit der Template Schreibweise in C++ und der allgemeinen Schreibweise in anderen Sprachen wie C#, Java, D und vielen weiteren.

Dies ist ein Trade-off welcher die häufiger genutzte Art der Deklaration vereinfacht und die seltener genutzte erschwert.

3.3.2.3 Pointer auf Arrays (Arithmetik auf Pointern oder auch nicht)

Des weiteren wird eine neue Schreibweise für Pointer auf Arrays eingeführt:

```
var int[*] a;
var int[*] a; // besser???, nein, weil diese syntax schon belegt ist?!
```

Damit lassen sich möglicherweise schon in Myll Fehler diagnostizieren wenn ein Pointer auf ein Skalar so verwendet wird wie ein Pointer auf ein Array z.B.:

```

var int    skalar;
var int[10] ary;
var int*   ptr_to_skalar = &skalar;
var int[*] ptr_to_ary    = &ary; // Expliziter Decay des Arrays zu einem Array-Pointer
var int[]* ptr_to_ary    = &ary; // Expliziter Decay des Arrays zu einem Array-Pointer, besser???
ptr_auf_skalar++;             // Compiletime Error: Pointer auf Skalare verbieten Arithmetik
ptr_auf_ary++;                // Funktionierte wie erwartet

```

Auch wird der implizite Decay??? von Arrays zu Pointern verhindert, will man, dass ein Array zum Pointer wird, nutzt man die identische Syntax um ein Pointer auf ein Skalar zu bekommen.

3.3.2.4 Die Kombination von Pointern und Arrays

Sofern man Variablen immer jeweils einzeln erzeugt oder man dem sich mit dem mentalem Modell aus [KR-CPL] anfreunden konnte, aber man dennoch nicht auf etwas Problematik verzichten will, dann kann man die beiden zuvor genannten Fälle kombinieren:

```
int *c[4];
```

Man weiß hier zwar das `*c[4]` ein `int` ist, aber der Teil mit dem `int` ist auch das leicht Verständlichste.

Ob dies ein Pointer auf ein Array mit 4 `int` ist oder ob es ein Array aus 4 Pointern auf `int` ist ist nicht direkt ersichtlich. Dafür muss man wissen welche Operation zuerst ausgeführt wird, die Dereferenzierung oder das Subskript. Und es stellt sich noch eine weitere Frage: Wenn es das Eine ist, wie bekommt man es dazu das Andere zu sein?

(Auflösung: Es ist ein Array mit 4 Pointern auf `int`. Die Umkehr wird durch Klammersetzung im Typen erreicht: `int (*c)[4];`)

Alternativ dazu sind die C++11 Äquivalente zwar Schreibintensiv, aber eindeutig und direkt verständlich:

```

unique_ptr<array<int,4>> uptr_to_ary;
array<unique_ptr<int>,4> ary_of_uptr;

```

Lösung

Die Kombination aus Pointern und Arrays ist durch die bereits genannten Änderungen schon abgedeckt und die Unklarheit der Reihenfolge, welche in C & C++ vorhanden war, ist nicht mehr vorhanden. Es gibt:

```

var int[4]* ptr_to_ary_of_4; // Ein Pointer auf ein Array aus 4 int
var int*[4] ary_of_4_ptrs;  // Ein Array aus 4 Pointern auf int

```

Dadurch besteht keine Notwendigkeit Klammern zusetzen und es ist auch nicht so Schreibintensiv wie die Template-Syntax.

3.3.2.5 Funktionspointer

Ein besonders schwer zu lesen und Schreibender Teil der C & C++ Syntax ist die der Funktionspointer. Hier 2 Beispiele:

```

int f(int a,int b) {...} // Definition der Funktion f mit zwei int Parametern und int Rückgabe
int ff(int(*f)(int,int)) {...} // Definition der Funktion ff die einen Parameter vom Typ von f hat
int(*fp)(int,int) = f; // Pointer auf eine Funktion fp vom gleichen Typ wie f, die auf f zeigt
void(*ffp)(int*)(int,int); // Pointer auf eine Funktion ffp vom gleichen Typ wie ff

```

Das Problematische hier sind die Funktionspointer deren Namen inmitten der Deklaration *versteckt* ist, welches sich noch undurchsichtiger bei verschachtelten Funktionspointern auswirkt, da diese Namenlos nur durch ein `"(*)"` identifiziert werden.

Die dritte Zeile kann auf den ersten Blick für eine *Functional Cast Expression* oder eine Deklaration eines Pointers auf ein *int* gehalten werden; vergleiche mit der Auflösung des vorangegangenen Punktes: `int(*c)[4];`.

Lösung

In Myll erkennt am linken Rand der Deklaration sehr schnell ob etwas eine Funktion, ein Funktionspointer oder etwas anderes ist, die Namen sind außerdem leicht auffindbar an der rechten Seite der Deklaration. Auch verschachtelte Funktionspointer sind nicht schwerer lesbar als Funktionspointer im Allgemeinen.

Das identische Beispiel von 2.2.2.5 in Myll:

```
func f(int a,int b)->int {...}
func ff(func(int,int)->int f)->int {...}
var func(int,int)->int fp = f;
var func(func(int,int)->int)->void ffp;
```

3.3.3 Schlechtes Standardverhalten

3.3.3.1 Ein Parameter Konstruktor

Ein Konstruktor, welcher einen einzelnen Parameter übernimmt, erfüllt nicht nur die Rolle eines Konstruktors mit einem Parameter, sondern er exponiert damit auch die Funktionalität der impliziten Konvertierung des Parametertypen zum Klassentypen. Dieses Default-Verhalten bringt den nicht-Profi in die Situation, Funktionalität bereitzustellen ohne davon zu wissen.

```
class C {
    C(int) {...}
    explicit C(float) {...}
};
C c1 = C(1);
C c2 = 2;
C c3 = C(3.0f);
C c4 = 4.0f;           // Compiletime Error
```

Der Entwickler wollte *c1* und bekommt *c2* ohne jede Warnung hinzu. Gleichfalls funktioniert *c3*, nur *c4* wird mit einem Compiletime Fehler verhindert.

Lösung

Die Lösung zum Konstruktor, welcher einen einzelnen Parameter übernimmt, ist die Umkehrung der Bereitstellung der impliziten Konvertierung. Sollte diese Konvertierung gewünscht sein, muss sie explizit erwähnt werden.

Auch Jonathan Müller, C++ Library Developer und Konferenzsprecher betrachtete dieses Problem in einem Blogpost und Urteilte [JM-EA]:

"As with most defaults, this default is wrong. Constructors should be explicit by default and have an implicit keyword for the opposite."

Dieses Beispiel in Myll weist die gleiche Bereitstellung wie im C++ Beispiel auf:

```
class C
{
    [implicit] ctor(int) {...}
    ctor(float) {...}
}
```

```
var C c1 = C(1);    // OK
var C c2 = 2;      // OK, man wollte explizit die implizierte Konvertierung von int erlauben
var C c3 = C(3.0f); // OK
var C c4 = 4.0f;   // Error
```

Der Entwickler bekommt ein Funktionierendes *c1* und wie angefordert auch *c2*. Gleichfalls funktioniert *c3*, jedoch wird *c4* mit einem Compiletime Fehler verhindert.

3.3.3.2 Private Ableitung

Ein wenig Text wäre hier wohl nicht schlecht???

```
class D : C {...};           // D leitet private von C ab, was keinen Sinn ergibt...
class D { C base; ... };     // ...sollte man dies gewollt haben, kann man es auch so schreiben
class D : public C {...};    // Dies ist der Fall den man üblicherweise will
```

Lösung

```
class D : C {...}           // Dies ist der Fall den man üblicherweise will
class D : private C {...}   // Wenn man will kann D private von C ableiten
```

3.3.3.3 Alles kann Exceptions werfen

Funktionen sind, wieder aufgrund von Kompatibilitätsgründen, auch ohne explizite Auszeichnung so definiert das sie Exceptions werfen können. Das führt dazu das man in Exception-Armen Projekten jede Funktion mit `noexcept` auszeichnen muss.

Ein Beispiel aus dem .NET Framework 1.1 in C#. Dort gibt es ein *Dictionary* (äquivalent zu *std::map*), welches Key/Value Paare speichert. Die Abfrage ob ein Key vorhanden ist wurde anfangs so gelehrt:

```
try {
    myDict[keyThatMightNotBeThere];
} catch(KeyNotFoundException) {
    // was passiert wenn der Key nicht vorhanden war
}
```

Sollte `myDict` erst über die Zeit wachsen ist das kein Ausnahmefall, es ist etwas was ständig passiert, für jeden neuen Key welcher dem *Dictionary* hinzugefügt wird. Die Performance war Verglichen mit der folgenden Exception-losen Alternative unterirdisch.

```
if(myDict.ContainsKey(keyThatMightNotBeThere)) {
    myDict[keyThatMightNotBeThere];
} else {
    // was passiert wenn der Key nicht vorhanden war
}
```

Obwohl hier im Falle des Vorhandenseins zwei mal nach dem Key gesucht wird, ist dies viel schneller als die `try/catch`-Variante.

In Version 2.0 von .NET Framework wurde `bool myDict.TryGetValue(value, out value)` eingeführt, es gibt zurück ob der Key gefunden wurde und legt den gefundenen Wert in `value` ab, alles mit nur einer Suche.

```
if(myDict.TryGetValue(keyThatMightNotBeThere, out val)) {
    val; // irgendwas mit val machen
} else {
    // was passiert wenn der Key nicht vorhanden war
}
```

So ist in C# seit Jahren die präferierte Nutzung von *Dictionary*.

Ein Zitat von Tim Sweeney von Epic Games im Januar 2020 auf Twitter:

"Just eliminated C++ exceptions in a complicated multithreaded library. It's 15% faster despite all performance sensitive functions previously being declared noexcept, with no obvious explanation."

Sinngemäß ins Deutsche: "Grade in einer komplexen Multithreaded Bibliothek C++ Exceptions eliminiert. Sie ist nun 15% schneller obwohl zuvor alle Funktionen auf dem kritischen Pfad noexcept markiert waren. Ich habe keine logische Erklärung dafür"

Für uns als Spieleentwickler sind 15% weniger Performance in Code welcher lediglich Exceptions bereitstellt und nicht einmal nutzt ein extremer Verlust. Diese Menge will kein aktuelles Spiel opfern.

Lösung

Exceptions sollen Ausnahmen sein, nicht die Regel. Nur wenn etwas *unerwartetes* passiert sollen sie auftreten. Etwa wenn jemand von Hand die Datei löscht welches das Programm mitbrachte und die eigentlich immer da sein sollte. Oder wenn die Netzwerkverbindung mitten im senden des Paketes abbricht.

Ich denke es hat sich erst nach und nach herausgestellt welcher Grad der Benutzung eines Sprachfeatures sinnvoll ist oder auch nicht. Das Problem welches die C++ zusätzlich hat, ist das dessen Exceptions Heap-Alloziert sind, obwohl sie es wohl technisch gar nicht sein müssten, der Standard schreibt es nur vor.

Ähnlich zu dieser Entwicklung in C# erreicht die berechtigte Kritik an der Überbenutzung von Exceptions das C++ Standardkomitee. Es wird darüber nachgedacht in weniger Funktionalität der Standardbibliothek Exceptions zu nutzen und auch ein neues invertiertes Keyword Namens `throws` einzuführen. Siehe diese Präsentation von Herb Sutter: <https://youtu.be/ARYP83yNAWk?t=2679>

Das, auch für zukünftige C++ Standards angedachte, Keyword `throws` ist in Myll nötig um in einer Funktion Exceptions nutzen zu können. Dieses Keyword kann in Myll auch einmalig einer gesamten Klasse zugewiesen werden und gilt dann für alle Member. Es bietet schon heute die Umkehr, welche sich im C++ Standard andeutet.

3.3.3.4 Switch fällt von selbst

Als *Labels* und *goto* noch ein oft genutzter Teil von Programmiersprachen war, konnte man noch einfach verstehen was an dem folgenden Beispiel falsch läuft. Da aber Heutzutage, bis auf das switch/case Statement, alle Kontrollstrukturen durch geschweifte Klammern begrenzt sind, tritt der hier gezeigte Fehler nicht selten auf.

```
switch( a ) {
    case 1:
        do_this();
    case 2:
        or_do_that();
        break;
    case 3:
    case 4:
        two_cases_want_the_same();
        break;
}
```

Hier wird im Falle das `a = 1` ist `do_this` ausgeführt und nach dessen ordnungsgemäßer Beendigung auch noch `or_do_that`. In den Fällen `a = 3` oder `a = 4` erwarten wir jedoch den *fallthrough* (das durchfallen) von `case 3` nach `case 4`.

C++ bietet schon die Möglichkeit der Auszeichnung das ein *fallthrough* erwünscht war, dieser wird aber nicht Erzwungen, siehe: <https://en.cppreference.com/w/cpp/language/attributes/fallthrough>

Lösung

Entweder `break` oder `fall` sind die letzten Statements in jedem `case`. Dadurch ist weder der eine noch der andere Fall des Vergessens möglich. Diese Sprache möchte einem nicht mühsam angewöhnte positive Verhaltensweisen abgewöhnen.

```
switch( a ) {
  case 1:
    do_this();
    fall;                // fällt in den darunterliegenden case, compile error ohne fall
  case 2:
    or_do_that();
    break;
  case 3, 4:             // wenn mehrere cases genau das gleiche wollen geht es auch so
    two_cases_want_the_same();
}
```

Oft sieht man auch das ein `case` seinen kompletten Inhalt in einen Block einbettet, um Beispielsweise lokale Variablen zu unterstützen (diese machen Probleme wenn sie nicht in einem Block eingebettet sind). Dieser Spezialfall könnte auf den Doppelpunkt verzichten.

Man könnte auch pro `switch` dessen default Verhalten entweder auf `break` oder `fall` Stellen, damit reduziert man Schreibarbeit, wenn stets das gleiche gewünscht ist (und kann mit `[default=fall]` den C++ Zustand wieder herstellen). Auch eine Auszeichnung das ein `switch` erschöpfend sein muss, also alle Fälle abdeckt, wäre insbesondere für Enums denkbar. Dies könnte Programmierfehler durch unbehandelte Fälle verhindern.

3.3.4 Inkonsistente Syntax

3.3.4.1 Funktions- & Methodendeklarationen

In C++ gibt es zahlreiche unterschiedliche Möglichkeiten Funktionen und Methoden zu definieren.

```
float pow(float b, int e)      {...} // nicht optimal weil fest vorgegebener Typ
auto pow(auto b, int e) -> float {...} // ab C++11, auto vorn überflüssig, aber syntaktisch nötig
auto pow(auto b, int e)      {...} // ab C++14
template <typename T>
T pow(T b, int e)             {...} // Analog auch auto Rückgabotyp, auch anhängend möglich
[](auto b, int e) -> auto      {...} // Lambda-Syntax, kein Rückgabotyp vorweg
```

Die Spezifikation der Typen von Parameter und Rückgabe ist in jedem aufgeführten Fall unterschiedlich, besonders die Positionierung des Rückgabetylen variiert stark. Die mit C++11 eingeführte Lambda-Syntax besitzt gar keine Möglichkeit eines vorangestellten Rückgabetylen mehr.

Lösung

So sehen die identischen Deklarationen in Myll aus.

```
func pow(float b, int e) -> float {...} // nicht optimal weil fest vorgegebener Typ
func pow(auto b, int e) -> auto  {...} // Rückgabotyp
func pow(auto b, int e)         {...} // das gleiche wie die Zeile davor
func pow<T>(T b, int e) -> T     {...} // Auch auto Rückgabotyp möglich
func(auto b, int e) -> auto      {...} // Lambda-Syntax, fast identisch zur Funktions-Syntax

proc do_something()              {...} // Momentan Synonym zu func
method draw()                   {...} // Innerhalb von Klassen Synonym zu func und proc
```

Funktionen und Methoden aller Art werden mit dem neuen Keyword `func` eingeleitet, die Rückgabetylen werden nur Folgend angegeben. In der Evolution von C++ wurde der folgende Rückgabotyp mit C++11 eingeführt und gewann seit dem immer mehr an Bedeutung und Funktionalität. Die Template-Syntax

wurde für simple Fälle wie diesen hier entschlackt, mehr dazu in separatem Kapitel. Lambdas sind nun leichter Identifizierbar und bis auf ihre Namenlosigkeit identisch in der Schreibweise zu Funktionen. Die Nutzung der Keywords *proc* und *method* ist auch möglich, *proc* ist momentan komplett Synonym zu *func*, *method* hingegen nur im Klassenkontext nutzbar.

Analog zu *const* Methoden, welche eine Veränderung des assoziierten Objektes verbieten, gibt es die Möglichkeit jede Funktion als *pure* zu markieren, welche die Veränderung von jeglichem globalen Zustand verbietet.

3.3.4.2 Notwendigkeit des Semikolons

Manchmal braucht man ein Semikolon nach Sprachkonstrukten und ein andermal braucht man es bei sehr ähnlichen Konstrukten nicht.

```
class E {...};
do {...} while(...);

namespace A {...}
while(...) {...}
if(...) {...}
```

Lösung

Ohne die Notwendigkeit, wie in C, für alles via *typedef* einen nutzbaren Alias zu erzeugen, fällt auch die Notwendigkeit des Semikolons nach Klassen, Strukturen, Unions und Enumerationen weg. Verloren ist die Möglichkeit der direkten Erzeugung von Variablen des neuen Typs.

```
class E {...}
do {...} while(...)
```

3.3.4.3 Verteilung der Attribute

In C++ sind die Attribute einer Funktion / Methode chaotisch um den Kern der Deklaration verteilt. Gegeben sei dieses Beispiel, zur besseren Lesbarkeit nach logischen Gruppen umgebrochen, formatiert und erklärt:

```
[[noreturn]] virtual const float const foo() const override final noexcept {...}

[[noreturn]]           // Attribut welches angibt das die Funktion nicht zurückkehren wird
virtual               // Methode ist virtuell überschreibbar
const float const      // Rückgabotyp
foo()                 // Methodennamen und leere Parameterliste
const override final noexcept // Methode verändert this nicht, ist überschrieben, nicht weiter...
{...}                 // ...überschreibbar und wirft keine Exceptions
```

Man beachte das *const* dreimal vorkommt, wobei sich die ersten zwei Vorkommen (redundant) auf den Rückgabeparameter, das dritte auf die Methode selbst bezieht. Man betrachte das *virtual*, *override* und *final* verwandte Terme sind (die innerhalb einer Definition wie hier aufgeführt nicht gemeinsam verwendet werden) aber dennoch an unterschiedlichen Positionen geschrieben werden.

Diese wahllos wirkende Verteilung der Attribute ist der Evolution von C++ geschuldet, *virtual* ist ein Keyword aus der Entstehung von C++, *override* und *final* sind relativ neu und lediglich *identifiers with special meaning*, welche nur in einem speziellen Kontext als Keyword dienen. In anderen Kontexten können sie etwa als Typ oder Variablenname verwendet werden, aus diesem Grund können sie nicht an der gleichen Position wie *virtual* geschrieben werden.

??? Standard zitieren

Lösung

Das gleiche Beispiel auch hier nach logischen Gruppen Formatiert:

```
[noreturn, pure, virtual, override, final, noexcept] func foo() -> const float {...}

[noreturn, pure, virtual, override, final, noexcept] // Alles was sich auf die Methode bezieht
func foo() -> const float                               // Keyword, Name und Rückgabotyp mit west-const
{...}
```

Myll gruppiert Attribute in Eckigen Klammern vor dem jeweiligen Konstrukt, dadurch sind zukünftige Erweiterungen einfach möglich, ohne das es Kollisionen mit bestehendem Code & Syntax gibt. *West const* ist hier die einzig Korrekte Schreibweise um etwas *const* zu machen.

3.3.4.4 Implizite Konvertierungen

In C++ sind viele unsichere implizite Konvertierungen möglich. Dies Betrifft insbesondere numerische Datentypen.

```
char a = 17, b = 4;
char c = a + b;
int d = 707, e = 1337;
char f = d + e;           // Konvertierungen zu kleineren Typen erzeugen keine Warnung
static_assert( -1 < 1u ); // Dies hier schlägt fehl, weil -1 implizit zu unsigned konvertiert wird
int * ip = 0;             // 0 ist implizit zu jeglichem Pointer konvertierbar...
int * jp = NULL;          // ...das ist so, weil das NULL Makro oft einfach als 0 definiert ist
if( a ) {...}             // Jegliche numerische Datentypen sind implizit zu bool konvertierbar,...
if( ip ) {...}            // ...was auch für Pointer gilt
int * sp = new int( 8472 ); // sp ist ein Pointer auf ein einzelnes int ...
sp++;                     // ... dennoch erlaubt es Pointer-Arithmetik ...
sp[23] = 9;               // ... und den Aufruf des Subscript-Operators
void * vp = sp;           // Alle Pointer sind implizit zu void Pointern konvertierbar
```

Das die Addition zweier *int* sich sehr Wahrscheinlich nicht vollständig in einem *char* speichern lässt, ist kein Zustand über den sich C++ beschwert. Noch absurder wird es wenn -1 nicht mehr kleiner als 1 sein soll. Das liegt daran dass für diesen Vergleich von *signed* und *unsigned int* beide in *unsigned int* konvertiert werden. Diese Konvertierung findet *höchst performant* statt, was bedeutet dass das -1 *signed int* einfach als *unsigned int* reinterpretiert wird und dann den Wert 4294967295 hat, welcher nicht kleiner als 1 ist.

Die Erklärung dafür ist das in C++ für alle Ganzzahlen, welche einen geringeren Wertebereich als *int* haben, bei arithmetischen Operationen eine Typenerweiterung durchgeführt wird. Das heißt das sie auf die Größe eines *int* erweitert werden, sollte jedoch der Wertebereich von *int* nicht ausreichen, wird alternativ zu *unsigned int* konvertiert. Dies passiert auch wenn der Wertebereich von *unsigned int* auch nicht ausreichend ist.

Diese Funktionalität wird **Integer Promotion** genannt.

Um einem Pointer mit dem besonderen leeren Wert zu initialisieren wird ihm entweder *0*, *NULL* oder *nullptr* zugewiesen, wobei aber nur der letztgenannte vom Typ Pointer ist.

Wird ein einzelnes Element als Pointer gespeichert, lässt sich auf diesen Pointer dennoch Pointer-Arithmetik und der Subscript-Operator anwenden, man könnte sagen ein *nicht arithmetischer Pointer* wird implizit in einen *arithmetischen Pointer* konvertiert.

Mögliche Lösung

Das man für die Berechnung von `c = a + b` keine Warnung erhält ist nachvollziehbar. Da C++ aber auch in diesem Fall eine Integer Promotion vornimmt ist das Resultat aus `a + b` vom Typen *int*. Eine Warnung in diesem Fall wäre ein *false positive* (eine Warnung welche keine hätte sein dürfen). Wenn ich also hierfür keine Warnung erhalten will, dann kann ich auch für `f = d + e` keine Warnung erwarten, weil das Ergebnis von `d + e` ebenso *int* ist.

Eine mögliche Lösung des Problems ist *Anzunehmen* das keine Integer Promotion stattfindet und dann generell eine strengere Typenüberprüfung anzusetzen. Hier würde dies bedeuten das `b + c` ein *char* ergeben würden, welches mit dem empfangenden Typen *char* übereinstimmt. Im Falle der Zuweisung von `f = d + e` würde nun ein *int* in ein *char* gespeichert werden, welches dank der Elimination des *false positive* nun eine Warnung erzeugen kann.

Unberührt dessen findet effektiv natürlich die Integer Promotion weiterhin statt, denn diese ist hilfreich um etwa dieses Beispiel hier vom Überlauf abzuhalten.

```
var char a = 28, b = 16;
var char c = (a * a) / b;
```

Ein Pointer kann ???

3.3.5 Problematische Reihenfolge der Syntax

Expressions (nicht nur Operatoren, so wie der Name vermuten lässt) sind im aktuellen C++ in 17 Ebenen einer Vorrangs-Tabelle eingeteilt.

Tabelle 3.1: Operator Precedence Table

Precedence	Operator	Precedence	Operator
1	::	7	e << e, e >> e
2	e++, e-- e() e[] e.e e->e typeid() const_cast dynamic_cast reinterpret_cast static_cast	8	e <==> e
3	++e, --e +e, -e !e, ~e, *e, &e (type) sizeof new, new[] delete, delete[]	9	e < e, e <= e e > e, e >= e
4	e.*e e->*e	10	e == e, e != e
5	e * e e / e e % e	11	e & e
6	e + e e - e	12	e ^ e
		13	e e
		14	e && e
		15	e e
		16	e ? e : e e = e e += e e -= e e *= e e /= e e %= e e <== e e >== e e &= e e ^= e e = e throw
		17	e, e

3.3.5.1 Unerwartete Reihenfolge der Operatoren

Manche dieser Ebenen sind in einer Reihenfolge, in welcher man sie nicht vermuten würde. Die Kommentare zeigen mittels Klammerung welcher Teil zuerst evaluiert wird.

```
cout << 1 & 2;           // (cout << 1) & 2;
if( a ^ b == 2 )      ... // if( a ^ (b == 2) ) ...
cout << a == b;       // (cout << a) == b;
```

Die Einordnung in der Vorrangs-Tabelle der Bitweisen *Und* und *Oder* Operation geht auf BCPL zurück. Dort erfüllte der Operator "&" kontextsensitiv entweder die Funktionalität der Konjunktion (logischem *Und*) oder des bitweisen *Und*, analog traf das selbe auf den Operator "|" und *Oder* zu. In der Entwicklung von C wurde die Funktionalität in & und && aufgespalten.

Quelle: <https://www.lysator.liu.se/c/clive-on-bcpl.html#operators>
https://en.wikipedia.org/wiki/Operators_in_C_and_C%2B%2B#Criticism_of_bitwise_and_equality_operators_precedence

<https://www.bell-labs.com/usr/dmr/www/chist.html>

Their tardy introduction explains an infelicity of C's precedence rules. In B one writes

```
if (a == b & c) ...
```

to check whether a equals b and c is non-zero; in such a conditional expression it is better that & have lower precedence than ==. In converting from B to C, one wants to replace & by && in such a statement; to make the conversion less painful, we decided to keep the precedence of the & operator the same relative to ==, and merely split the precedence of && slightly from &. Today, it seems that it would have been preferable to move the relative precedences of & and ==, and thereby simplify a common C idiom: to test a masked value against another value, one must write

```
if ((a & mask) == b) ...
```

where the inner parentheses are required but easily forgotten.

Die Notwendigkeit B Programme nach C umzusetzen ist nicht mehr gegeben, dennoch schleppt C++ immer noch diesen Ballast mit sich herum.

Lösung

In Java und C# sind die Operatoren &, | und ^ in ähnlichem Vorrang wie in C++ und seinen Vorgängern einsortiert und diese verhalten sich je nach booleschen oder ganzzahligem Kontext analog zu C++.

Die Sprachen Rust und Ruby hingegen sortieren die Operatoren &, | und ^ direkt über die Vergleichsoperatoren ein und eliminieren die in Kapitel 2.2.5.1 angesprochene Problematik.

Myll ordnet die booleschen Operatoren in die Ebenen der Punkt- und Strichrechnung ein: Operator & in die Ebene der Punktrechnung, | und ^ in die Ebene der Strichrechnung.

Die Idee hierzu kam durch das Projekt Bitwise von Per Vognsen welcher in seiner Programmiersprache **Ion** die gleiche Änderung der Operator Precedence umgesetzt hat.

Welche Sprachen machen dies noch so: GO ???

Der letzte aufgeführte Problemfall wurde nicht Behandelt damit dies hier noch so funktioniert:

```
if( a << 1 == 23 ) // if( (a << 1) == 23 )
```

Siehe auch 4.1.4.3. what??? leider Pointer auf umbenanntes Kapitel

3.3.5.2 Pre- und Suffix Operationen

Ähnlich zur Variablendeklaration von Pointern in Kombination mit Arrays bietet auch das Zusammenspiel von Pre- und Suffix Operationen oft eine Ambiguität mit sich, welche Klammersetzung erfordert.

Verwirrend ist auch die Nutzung der gleichen Symbole für unterschiedliche Arten von Operationen besonders bei: $\&$, $*$, $+$, $-$

Diese werden zum einen für Unäre Prefix- sowie Binäre Operationen verwendet, als auch in doppelter Erwähnung als Prefix, Suffix und Binäre Operation.

[illegible]

Die Ziffern in der untersten Zeile sagen aus in welcher Reihenfolge die Teile dieser Expression ausgewertet werden, 1 ist zuerst, dann aufsteigend bis 8. In fett sind die Stellen markiert welche einen Unterschied von mehr als Eins zu ihrem Nachbarn haben, hier zerbricht der Lesefluss oder besser er *sollte* zerbrechen. Wenn er nicht zerbricht heißt dass das man wichtige Informationen übersieht. Im obigem Beispiel sind **fünf** dieser Stellen.

Lösung

Dieser Fall bleibt so, Pech:

```
(*a)->draw(); // Auch bei mehrfach Dereferenzierung muss geklammert werden
```

Wären alle Unären Operationen Suffix Operationen dann könnte das Beispiel so aussehen

[illegible]

In diesem Beispiel sind nur noch **zwei** Stellen welche den Lesefluss unterbrechen.

Eine generelle Umstellung geht aber leider nicht, weil z.B. ein Suffix-`*` mit dem Multiplikations-Operator kollidiert, Suffix-`++` schon belegt ist und `-e` für den Vorzeichenwechsel einfach besser aussieht als `e-`.

Es gibt in C++ schon die Möglichkeit z.B. `e.operator!()` aufzurufen, was aber nicht auf primitiven Typen funktioniert und außerdem so schreibintensiv ist, das es entweder deshalb nicht genutzt wird oder schon durch seine Präsenz allein den Lesefluss beeinträchtigt. In der gleichen Art könnte man zusätzlich zur normalen Schreibweise z.B. `e.not`, `e.neg`, `e.addr`, `e.preinc` als Pseudomember auf die Relevanten Typen einführen, was jedoch eine Explosion an Keywords mit sich bringen würde.

Die Lösung für Myll ist das es aktuell keine zufriedenstellende Lösung gibt.

3.3.5.3 Zuweisungen und Throw Expressions

```
if( a = 1 )           // Fälschliche Zuweisung
??? throw bspl
```

Lösung

Zuweisungen und *throw* sind keine Expressions mehr. Dadurch sind die genannten Beispiele nicht mehr valider Code und müssen in einfacher durchschaubaren Code umgeschrieben werden. In den seltenen Fällen wo eine Zuweisung eine Expression war muss sie nun einzeln als Statement geschrieben werden. C# und ??? löst es genau so.

3.3.6 Schwer durchschaubare automatische Verhaltensweisen

3.3.6.1 Special Member Functions

Es gibt sechs spezielle an eine Klasse gebundene Funktionen in C++ und diese werden unter jeweils unterschiedlichen Bedingungen automatisch generiert.

Siehe: https://en.wikipedia.org/wiki/Special_member_functions

- Default Constructor
- Destructor
- Copy Constructor
- Copy Assignment Operator
- Move Constructor
- Move Assignment Operator

Problematisch ist dies in vielerlei Hinsicht. Zuerst muss man sie überhaupt erkennen, um sie mit der notwendigen Sorgfalt zu behandeln, denn es kann abgesehen von diesen speziellen Typen beliebig viele normale Konstruktoren und Zuweisungsoperatoren geben. Drei von den hier Aufgelisteten sind Konstruktoren, zwei davon erwarten einen einzelnen Parameter mit besonderem Typ. Die speziellen Zuweisungsoperatoren erwarten diesen gleichen Parameter Typen.

Sobald man Einen davon selbst implementiert, fallen automatisch einige der Anderen weg. Dies ist schwer durchschaubar und wirft im Normalfall noch zusätzliche Probleme auf, so das die Verhaltensregeln *Rule-of-Three*, *Rule-of-Five* und *Rule-of-Zero* erdacht wurden. Diese geben vor, das wenn man eine der speziellen Funktionen implementiert (mit Ausnahme des *Default Constructors*), das man in dem Zuge auch einige der anderen Implementieren soll. Anders der Fall des *Rule-of-Zero*, welcher vorgibt das man komplett auf die Implementierung der speziellen Funktionen verzichten sollte, wiederum mit Ausnahme des *Default Constructors*. Siehe auch [JM-SM].

Diese Regeln existieren auf dem Papier, aber leider nicht in der Sprache, somit ist die Validation von deren Befolgung auch komplett an den Benutzer übertragen.

Standard from 10.2/2 ???

Zusammenfassung

- Es ist schwer zu erkennen ob und welche Probleme vorhanden sind
- Der Compiler validiert nichts davon

Lösung

```
[rule_of_n=5] // Dies wirft einen Fehler beim Kompilieren: Es fehlt dtor und move=
class Test {
    move_ctor(other) {...} // Der Leser sieht direkt das es sich um einen Move Constructor handelt
    copy_ctor(other) {...} // Der Leser sieht direkt das es sich um einen Copy Constructor handelt
    copy=(other) {...} // Der Leser sieht direkt das es sich um ein Copy Assignment handelt
}
```

3.3.6.2 Overloading mit Vererbung, Shadowing

<https://www.geeksforgeeks.org/does-overloading-work-with-inheritance>

```
class Base {
    void f(int) {...};
}
class Derived : public Base {
    // using Base::f;      // diese Zeile würde das Problem beheben
    void f(double) {...};
}
int main() {
    Derived d;
    d.f(1337);           // Ruft Derived::f(double) auf und nicht das logischere Base:f(int)
}
```

Lösung

Inversion des Verhaltens, erwarte vom Benutzer das er explizit erwähnt das er Shadowing will.

```
class Derived : public Base {
    void f(double) {...};           // using Base::f; wird automatisch hinzugefügt
}
class OtherDerived : public Base {
    [shadow]
    void f(double) {...};           // using Base::f; wird nicht hinzugefügt
}
int main() {
    Derived d;
    d.f(1337);                     // Ruft Base::f(double) auf
    OtherDerived d2;
    d2.f(1338);                    // Ruft wie explizit angefordert Derived::f(double) auf
}
```

3.3.6.3 Zweierlei Enums

Es gibt zwei verschiedene Arten von Enums in C++. Das erste ist dass, welches schon seit C Zeiten vorhanden ist. Es hat den Nachteil das seine Member nicht nur unter dem Namen des Enums erreichbar sind, aus diesem Grund werden im Normalfall dessen Member mit dem Enumnamen präfigiert.

Die von C++ eingeführte Variante der Enums mit dem Zusatz `class` beseitigt das Problem der ungewollten Veröffentlichung aller Member in den aktuellen Namensraum, sie heißen *Scoped Enumeration*. Es wird allerdings auch die automatische Konvertierung in das zugrundeliegende Integer abgeschaltet. Das ist grundsätzlich etwas Gutes, hat aber auch negative Begleiterscheinungen, zum Beispiel funktionieren bitweise Operatoren erst nach expliziter Implementierung für den speziellen Enum Typ.

```
enum A { X = 1, Y = 2, Z = 4 };
enum B { X, Y, Z };               // die Membernamen X, Y, Z, kollidieren mit denen von A
enum C { C_X, C_Y, C_Z };         // Präfigierung mit dem Namen des Enums
A a = A::X | A::Y;                // Bitweise Operatoren funktionieren auf diesen Enums

enum class D { X = 1, Y = 2, Z = 4 };
enum class E { X, Y, Z };         // kollidiert nicht mit den Membern von D
D d = D::X | D::Y;                // funktioniert nicht so einfach, erst nach Definition von:
D operator|(D lhs, D rhs) {       // normalerweise mit static_cast und underlying_type
    return (D)((int)lhs|(int)rhs); // dies muss für alle bitweisen Operatoren wiederholt werden
}
```

Lösung

In Myll sind alle Enums *scoped*, sie schreiben sich durch das Keyword `enum`. Sollte man auf sie bitweise Operatoren anwenden wollen, zeichnet man einfach das Enum mit `[operators(bitwise)]` aus und man bekommt die Standardimplementierung dieser. Sollte man exponentielle Durchnummerierung der Elemente wünschen lässt sich dies mittels `[flags]` einschalten. Enums in C# haben ein `[Flags]` Attribut welches leider nur die Textuelle Ausgabe via `ToString()` verbessert, bitweise Operatoren hat in C# jedes Enum.

```
[operators(bitwise), flags]
enum A { X, Y, Z };           // automatisch X = 1, Y = 2, Z = 4
```

3.3.7 Schlechte & Schreibintensive Namensgebungen

C++ hat viel schlechte Namensgebung von C geerbt. Manche hat aber auch C++ selbst erfunden und die eigentlich gut lesbare von C verdrängt.

3.3.7.1 Datentypen

Es gibt die sehr schreibintensive Variante um eine *vorzeichenlose sehr sehr große Ganzzahl* zu definieren mittels *unsigned long long int*, das ist fast soviel zu schreiben wie es im Deutschen hier geschrieben steht.

Und es ist nicht einmal portabel verlässlich wie viel Bit an Präzision man von diesem Typen erwarten kann. Ein *short* hat üblicherweise 16 Bit, dürfte aber auch 32 haben. Ein *int* darf laut C++ Standard auch 16 Bit haben, hat üblicherweise 32 und es ist auch nicht verboten das es 48 oder 64 hat. Besonders prekär ist der Typ *long* denn er verhält sich auf aktuellen 64 Bit Betriebssystemen unterschiedlich, auf Windows nämlich als 32 Bit große Zahl, konträr dazu auf Linux, MacOS und UNIX mit 64 Bit.

Auf diesen Freiheiten möchte man keinen Serialisierer schreiben, welcher Daten in Binärform speichert. Deshalb gibt es in *cstdint* die Definition der Typen: `std::int8_t`, `std::int16_t`, `std::int32_t` und `std::int64_t` und die identischen *unsigned* Varianten mit dem Präfix `u`. Diese Datentypen sind optional, es heißt man kann sich nicht darauf verlassen das sie existieren. Außerdem verleiten sie Einsteiger dazu `using namespace std;` in ihren Programmen zu verwenden oder sie nutzen weiter die variablen Typen.

Lösung

Myll führt durchschaubare und aufgeräumte Namen für eingebaute Typen ein. Die meisten besitzen eine feste Bitbreite und die wenigen welche einen dynamische Größe haben sind streng limitiert in ihrer möglichen Bitbreite.

Vorzeichenbehaftete Ganzzahlen: `i8`, `i16`, `i32`, `i64`, `int`, `isize`

Vorzeichenlose Ganzzahlen: `u8`, `u16`, `u32`, `u64`, `uint`, `usize`

Gleitkommazahlen: `(f16,)` `f32`, `f64`, `f128`, `float`

Alle Typen welche Zahlen enthalten, haben genau die angegebene Bitbreite. Die Typen `int` und `uint` haben mindestens 32 Bit, nutzen aber möglicherweise mehr. Die Typen `isize` und `usize` entsprechen größtenteils `ptrdiff_t` und `size_t` und haben mindestens 32 Bit, aber immer genug Platz um einen normalen Pointer zu speichern. Der Typ `f16` ist optional (half precision floating point number) und `float` ist mindestens 32 Bit Breit, möglicherweise aber auch mehr.

Sollte man Daten Serialisieren wollen, um sie etwa abzuspeichern oder übers Netzwerk zu verschicken, wird dazu geraten dies nur mit den Datentypen von fester Breite umzusetzen. Es wäre sogar möglich die Nutzung der variablen Typen in Strukturdefinitionen welche der Serialisierung dienen zu verbieten.

3.3.7.2 C++ und C-Style Casts

C-Style Casts haben eine leicht les- und schreib-bare Form, leider hat ihre assoziierte Semantik dazu geführt das sie immer weniger in C++ genutzt werden. Die originale Semantik von C umsetzend, bieten die C-Style Casts in C++ die automatische Verhaltensweise das sie unterschiedliche `*_cast` Expressions kombinieren um den Typen umzuwandeln.

Siehe Auszug aus https://en.cppreference.com/w/cpp/language/explicit_cast

When the C-style cast expression is encountered, the compiler attempts to interpret it as the following cast expressions, in this order:

- a) `const_cast<new_type>(expression);`
- b) `static_cast<new_type>(expression)`, with extensions: pointer or reference to a derived class is additionally allowed to be cast to pointer or reference to unambiguous base class (and vice versa) even if the base class is inaccessible (that is, this cast ignores the private inheritance specifier). Same applies to casting pointer to member to pointer to member of unambiguous non-virtual base;
- c) `static_cast` (with extensions) followed by `const_cast`;
- d) `reinterpret_cast<new_type>(expression);`
- e) `reinterpret_cast` followed by `const_cast`.

The first choice that satisfies the requirements of the respective cast operator is selected, even if it cannot be compiled...

Aufgrund dieser ausgearteten Verhaltensweise verbannen sehr viele Projekte C-Style Casts komplett oder erlauben sie lediglich für Grundlegende Typen oder POD (Plain Old Data, salopp: alle Strukturen welche genau so auch in C vorkommen können). Man kann sagen Sie sind *quasi* ausgestorben.

Außerdem ist `std::move(val)` im Grunde nur ein Cast des enthaltenen `val` in einen *aus mir kann der Inhalt verschoben werden* Typ.

Lösung

In Myll fällt die ursprüngliche Verhaltensweise des C-Style Casts weg und `(Typ) val` ist **nur** noch ein `static_cast`, weil dieser am meisten verwendet wird. (Belegen???) Der Aufruf mit `(? Typ) val` ist ein `dynamic_cast` und `(! Typ) val` ist je nach gegebenem `Typ` entweder ein `const_cast`, ein `reinterpret_cast` oder beides. Sollte man hier eine Unterscheidung wünschen, ist natürlich die schreibintensive Variante (für alle Casts) weiterhin verfügbar.

Analog zu den grade beschriebenen Casts verhält sich auch `std::move`, ich kann mit dem Aufruf `(move) val` ein Objekt zu einem verschiebbaren Objekt *casten*.

3.3.7.3 Raw- & Smartpointer

Auch Raw-Pointer haben eine leicht les- und schreib-bare Form, leider hat ihre assoziierte Semantik dazu geführt das sie immer weniger in C++ genutzt werden.

Die Nutzung von **Rawpointern** wurde durch die Einführung von Smartpointern der Standardbibliothek, sowie von Boosts, als auch selbstgeschriebenen Intrusive-Smartpointern und Handles **floohh???** weit zurückgedrängt. Dies hat die relative Nutzungshäufigkeit von Rawpointern zurückgehen lassen. Raw-Pointer haben aber auch einen nicht zu unterschätzenden Vorteil, sie sind wesentlich übersichtlicher zu lesen und selbst in modernsten C++ Projekten muss jeder generell mit Rawpointern umgehen können.

Mehr zur Parameterübernahme von Smartpointern kann in [HS-SPP] nachgelesen werden.

Ausnahmsweise sind hier nicht die Keywords **fett** gedruckt sondern die Datentypen der Parameter, nur *const* ist hier *kursiv*, es wurde auch mehr Energie in eine lesbare Formatierung gelegt als Üblicherweise.

```
void composeGUI( Widget *parentElement, unique_ptr<const ScrollableWidget> &scrollElement,
weak_ptr<ButtonWidget> &buttonChild, unique_ptr<Widget> anotherChild ) {...}

void composeGUI(
```

```
const Widget      * parentElement,
unique_ptr<ScrollableWidget> & scrollElement,
weak_ptr<const ButtonWidget> & buttonChild,
unique_ptr<Widget>  anotherChild
) {...}
```

Man kann jetzt meinen dieses Beispiel ist *konstruiert*, na klar ist es das: und zwar um Probleme aufzuzeigen welche in echten Projekten mit Unübersichtlichen Parameterlisten auftauchen. Ich habe schon viel unübersichtlicheren Code (welcher Produktiv/Live/Ausgeliefert ist) gesehen als dieses Beispiel hier.

???Herb Sutter how pass smartpointers oder so

Verstecktes Verhalten welches hier leicht zu übersehen ist: Der Parameter scrollElement wird als Referenz auf einen Unique-Pointer übergeben, das bedeutet das die aufgerufene Funktion den Besitz übernehmen kann, aber nicht muss. Der Besitz des anderen Unique-Pointers anotherChild wird auf jeden Fall an die Funktion übergeben. Dieser wichtige Unterschied geht schnell unter weil der Text `unique_ptr` weit vom (nicht) vorhandenen `&` entfernt steht. Das `const` welches das ButtonWidget vor Veränderung schützt ist ebenso leichter zu übersehen als beim Raw-Pointer auf Widget.

Unübersichtlich wird es auch durch die sehr weite Entfernung des Zieltypen von der linken Seite.

Lösung

Smartpointer bekommen in Myll Syntactic Sugar verpasst. An den Positionen im Code, an denen Raw-pointer auftreten können kann ein `unique_ptr` sich als `*!` schreiben lassen, ein `shared_ptr` als `*+` und der `weak_ptr` als `*?`. Diese Symbole wurde Aufgrund ihrer Ähnlichkeit zu *Regular Expressions* gewählt, ein *Plus* zeigt *Mehrfachnutzung* an, ein *Fragezeichen* *Optionalität*. Sollte man nicht die Smartpointer der Standardbibliothek nutzen wollen, sondern etwa die von Boost oder seine eigenen, lässt sich dies leicht Umstellen.

Hier das Smartpointer Beispiel in Myll.

```
func composeGUI( Widget * parentElement, const ScrollableWidget *!& scrollElement, ButtonWidget *?&
buttonChild, Widget *! anotherChild ) -> void {...}

func composeGUI(
const Widget      * parentElement,
ScrollableWidget *!& scrollElement,
const ButtonWidget *?& buttonChild,
Widget           *! anotherChild
) {...}
```

Hier sind die Indirektionsebenen direkt beieinander, nicht an unterschiedlichen Stellen der Deklaration, Feinheiten wie die *erzwungene* oder die *mögliche* Übergabe des Besitzes entgehen einem nicht mehr so leicht. Das Keyword `const` welches sich auf das bezielte Element bezieht, ist stets ganz links abzulesen.

3.3.7.4 Funktionale Programmierung

Die Sprache C++ zählt Funktionale Programmierung zu seinen verfügbaren Programmierparadigmen. Schauen wir uns ein relativ simples Beispiel an und vergleichen dies mit der nicht-funktionalen Implementierung, als auch mit der Lösung welche andere Programmiersprachen bieten.

Als simples *Map & Reduce* Beispiel soll diese einfache Berechnung dienen. Eine Liste von Zahlen soll zuerst durch zwei dividiert und dann das Produkt gebildet werden. Die drei Zahlen im Container (4, 6, 14) werden zunächst in *map* jeweils durch 2 dividiert, dann wird in *reduce* das Produkt gebildet und ergibt die Zahl 42.

```
using namespace std;
```



```
// erste Variante
vector<int> c = {4, 6, 14};
transform(begin(c), end(c), begin(c), [](auto d){ return d / 2; });
const int fortyTwo = accumulate(begin(c), end(c), 1, multiplies<int>());

// zweite Variante, zeichen anzahl bestimmen???
vector<int> c = {4, 6, 14};
const int fortyTwo = transform_reduce(
    begin(c), end(c), 1, multiplies<int>(), [](auto d){ return d / 2; });
```

Die erste Variante verändert den Ursprungs-Container (was möglicherweise unerwünscht ist).

Die zweite Variante schreibt verwirrenderweise die Reduktion vor der Transformation.

In C# würde es äquivalent so aussehen, ohne störende Nebeneffekte:

```
var c = new List<int>{4, 6, 14};
int fortyTwo = c.Select( d => d / 2 ).Aggregate( (accu, e) => accu * e );
```

Ein simples *Map & Reduce* Beispiel in Ruby, auch ohne störende Nebeneffekte:

```
# erste Variante
c = [4, 6, 14]
fortyTwo = c.map{ |d| d / 2 }.reduce{ |accu, e| accu * e }

# zweite Variante
c = [4, 6, 14]
fortyTwo = c.map{ |d| d / 2 }.reduce( :* )
```

Wie sieht es mit der nicht funktionalen Implementierung in C++ aus? Die folgenden nicht-funktionalen Implementierungen in C++ sehen viel lesbarer aus, viel eher wie die funktionalen Implementierungen in C# und Ruby, da stellt man sich natürlich die Frage: Wieso schaffen es die Sprachen, welche schon viel länger Lambdas und Funktionale Bibliotheken haben als C++, dennoch in der funktionalen Implementierung lesbarer zu sein?

```
// erste Variante
vector<int> c = {4, 6, 14};
int accu = 1;
for (auto& d : c) { d /= 2; }
for (auto e : c) { accu *= e; }
const int fortyTwo = accu;

// zweite Variante (diese Variante cheatet ein wenig, weil sie beide Schritte direkt verschmelzt)
vector<int> c = {4, 6, 14};
int fortyTwo = 1;
for (auto d : c) { fortyTwo *= d / 2; }
```

Anmerkung zu den C++ Beispielen: Die funktionalen Implementierungen haben Vorteile gegenüber den nicht funktionalen Versionen. Zum einen kann zur potentiellen Steigerung der Performance `transform`, `reduce` und `transform_reduce` *parallel* sowie *nicht-sequenziell* ausgeführt werden. Zum anderen wird entweder `accu` nach außen hin sichtbar oder es wird mehrfach in `fortyTwo` geschrieben und es kann deshalb nicht als `const` markiert werden.

Potentielle Lösung (aktuell noch nicht in Myll implementiert)

Ruby hat die Methoden `map` und `reduce` auf seinen *Containern* `Array` und `Enumerable` definiert. Das äquivalent dazu wäre die C++ Container um diese zu erweitern, was das Ziel von Myll verfehlen würde (Keine Neu-Implementierung der C++ Klassen).

C# geht einen anderen Weg, es erlaubt statische Methoden durch *Syntactic Sugar* an Objekte zu binden, diese Funktionalität wird Erweiterungs-Methoden (Extension Methods) genannt. In einem Schritt des C# Compilers wird etwa der Aufruf von `c.Select(lambda)` in `Select(c, lambda)` übersetzt, dies passiert aber nur bei speziell markierten statischen Methoden.

Die Erweiterungs-Methoden von C# beschreiben einen Subset dessen was UFCS (Universal Function Call Syntax) erlaubt, nämlich bei jeder Funktion das erste Argument auch vor die Funktion zu ziehen und sie wie eine Methode aufzurufen. In UFCS wäre etwa auch `3.pow(3)` möglich, welches in C# nicht möglich ist, weil `pow()` nicht als Erweiterungs-Methode markiert ist. In D und in Nim gibt es UFCS.

Myll könnte entweder UFCS einführen, welches sich extrem auf die komplette Sprache und deren Nutzung auswirken könnte, oder es könnte den gleichen Weg wie C# gehen und nur selektiv und vorsichtig einzelne sinnvolle Funktionen an Objekte binden.

C++20 soll eine *Ranges* genannte Erweiterung in der Standard Bibliothek erhalten, welche möglicherweise den aktuellen Missstand mit der Funktionalen Programmierung behebt.

3.3.8 Keyword Wiederverwertung

C++ hat eine inflationäre Nutzung von immer wieder den gleichen Keywords in unterschiedlichen Kontexten. So kann **static** entweder die Veröffentlichung eines Identifiers über Translation Unit Grenzen verstecken, eine Variable innerhalb einer Funktion persistent machen oder innerhalb einer Klasse Variablen und Funktionen schaffen die global sind anstelle sich auf die Instanz zu beziehen. Logischerweise wird das Keyword nicht genutzt um *statische Bindung* im Gegensatz zur *dynamischen Bindung* (via *virtual*) bei Klassen zu signalisieren.

Das Keyword **const** kann angeben ob eine Variable nicht mehr als ein mal beschrieben werden darf oder ob eine Methode den Zustand des Objekts nicht ändert. Also ist const in keinem der Fälle zu verwechseln mit mathematischen Konstanten, welche immer den gleichen Wert haben; das Keyword *constexpr* erfüllt eher diese Eigenschaft in C++.

Lösung

```
class Test {
    [global,                // Hier kann auch static stehen, global ist nur sprechender
     readonly]             // Hier kann auch const stehen, readonly ist nur sprechender
    var i32 notInstance = 1337;
    [pure]                 // Hier kann auch const stehen, pure ist nur sprechender
    method man() {
        [persist]          // Hier kann auch static stehen, persist ist nur sprechender
        var f32 expensiveResult = expensiveCalc();
    }
}
```

Wer sich nicht an neue Keywords gewöhnen mag wird trotzdem glücklich, denn die alten funktionieren auch noch.

3.3.9 East Const und West Const

Gangkriminalität zwischen verfeindeten Clans ist hier nicht gemeint, es geht lediglich um die mögliche Positionierung des Keywords *const*. (Auch wenn es auch hier analoge Auseinandersetzungen wie zwischen der *geschweifte Klammer in neuer Zeile* mit der *geschweifte Klammer auf der gleichen Zeile* Fraktionen gibt)

```
const int * a;
int const * b;
int * const c;
```

Die Variablendeklaration von *a* und *b* sind identisch, das *const* bezieht sich auf das *int*. Bei der Deklaration von *c* hingegen bezieht sich das *const* auf den Pointer. Das heißt dass es an manchen Stellen möglich ist das *const* links von der zu attributierenden Stelle zu schreiben (west const) und an anderen rechts (east const).

Auch wenn es wesentlich häufiger der Fall ist *a* anstelle von *b* anzutreffen, gibt es in C++ nur eine Schreibweise die immer funktioniert, welches natürlich der unnatürlichere Fall des *east const* ist.

Lösung

Myll setzt konsequent ***west const*** ein, *east const* ist nicht valide.

3.3.10 Präprozessor

Der Präprozessor ist eine Schattenwelt welche teilweise ähnliche Funktionalität wie C++ selbst bereitstellt, nur leider mit einer komplett anderen Schreibweise, neuen Freiheiten sowie vielen Einschränkungen daher kommt.

Viel der Funktionalität welche ursprünglich nur von Präprozessor ausgeführt werden konnte, kann C++ schon selbst, teils sogar besser, erfüllen.

```
#define MAX(l,r) ((l)<(r)?(r):(l)) // Hat potentiell Nebeneffekte durch Mehrfachausführung
template <typename T>
T max(T l, T r) { return l<r?r:l; } // Keine Nebeneffekte und sogar Typsicher

#define ARY_SIZE 707
constexpr int ary_size = 707;
int ary[ary_size]; // Geht dank constexpr

#ifdef DEBUG
    log.print("only in debug mode");
#endif
if constexpr (DEBUG) { // Geht seit C++17, aber nicht außerhalb von Funktionskörpern
    log.print("only in debug mode");
}
#ifdef DEBUG // Geht nicht mit if constexpr
#include <debug.h>
#endif

#if !__cpp_noexcept // Funktionalitätsprüfung mit leerem Polyfill
#define noexcept // Aktuell nicht ersetzbar in C++
#endif
```

Lösung

Zur Lösung des Präprozessor *if* nutzt Myll Attribute. Alles was Myll nicht löst, kann in .cpp Dateien erledigt werden, der von Myll erzeugte Output ist ja Kooperativ mit *normaler* C++ Code kompatibel.

```
[Debug] // Führt das attributierte Statement nur im Debug Mode aus
log.print("only in debug mode");

[CT] // Erzwingt die Ausführung zur Compiletime, nicht nur bei if
if( someConstant ) ...

// Polyfills sollten nicht nötig sein
```

3.3.11 Goto

Das *goto* Statement und die Labels zu denen mittels *goto* gesprungen werden kann sind Heutzutage für die meisten Nutzungen verpönt. Eine der wenigen Ausnahmen dessen, ist der Fall das man mehr als eine Schleife verlassen will.

Lösung

Myll bietet die Möglichkeit bei *break* und *continue* Statements eine Zahl mitzugeben wie viele Schleifen verlassen oder fortgesetzt werden sollen. Hier ein Beispiel mit der neuen *loop* und *times* Schleife

```
loop {                                     // #1. Endlose Schleife
  do 10 times i {                         // #2. Zählschleife
    if( something[i] ) continue;          // macht bei #2 weiter, identisch mit continue 1;
    if( someother[i] ) break;             // macht bei #3 weiter, identisch mit break 1;
    if( somewhat[i] ) continue 2;        // macht bei #1 weiter
    if( someelse[i] ) break 2;           // macht bei #4 weiter
  }
  cout << "10x fertig\n";                 // #3
}
cout << "infloop fertig\n";              // #4
```

4 Implementierung

4.1 Umsetzung

Meine Arbeit mit TypeScript (welches das Arbeiten mit JavaScript erträglich macht???) und ähnlichen Projekten, welche eine Source to Source Übersetzung (Transpilation) durchführen, brachte mich dazu dieses Projekt mit dem gleichen Verfahren zu beginnen.

Durch Transpilation entgeht man vielen problematischen und arbeitsintensiven Aspekten wie:

- Erzeugung von Assembly / Maschinencode
- Optimierung von Assembly / Maschinencode
- Erzeugung eines ABI (Application Binary Interface, Binäre Repräsentation)
- Linking von Object Files
- Bereitstellung eines Debuggers

Es bietet einem die Möglichkeit eine Sprache zu bedienen, ohne sie schreiben zu müssen, sowie die gesamte Infrastruktur der Zielsprache nutzen zu können.

Auch andere aktuelle Entwicklungen nutzen Transpilation um schneller einsatzbereit zu werden, wie zum Beispiel Jonathan Blow's **Jai**, welches anfänglich zu C++ Code transpilierte und Per Vognsen's **Ion** welches immer noch zu C Code transpiliert. C++ selbst war in seinen Anfängen (**C with Classes**) als Transpiler zu C-Code umgesetzt.

Myll soll direkt mit C++ Code zusammenspielen, also einer Mixtur aus .cpp/.h und .myll Dateien, es soll die C++ Standardbibliothek, sowie C und C++ Bibliotheken nutzen können.

Die Umsetzung erfolgt in C# und nutzt Antlr4 als Lexer und Parser Generator.

4.1.1 C++ predigen, aber C# und Java trinken

Man sollte immer das beste Werkzeug für seine Arbeit nutzen. Aufgrund der Komplexität eine Sprache wie C++ zu Lexen und zu Parsen, bediene ich mich eines weitverbreiteten Parser-Generators und einer einfacher zu nutzenden Sprache. Die *Performance* welche bei dieser Arbeit im Vordergrund steht, ist die Umsetzung von möglichst viel Funktionalität im gegebenen Zeitfenster, dies lässt sich mit C# einfacher realisieren. ANTLR ist in Java geschrieben und erfordert lediglich für die Übersetzung der Grammatik ein installiertes JRE (Java Runtime Environment).

4.1.2 Alternative Umsetzungsmöglichkeiten

Im Kontext der Erstellung dieser Masterarbeit habe ich auch Alternativen zur Umsetzung betrachtet.

- Selbstgeschriebener Lexer/Parser
 - Nach einigen kleinen Prototypen war klar das eine Programmiersprache, welche annähernd die Komplexität von C++ aufweist, sich nicht in absehbarer Zeit manuell Lexen & Parsen lässt
 - Dieses Unterfangen wäre sicher Interessant gewesen, ist hier aber nicht der gewünschte Fokus
- C++ mit Boost Spirit X3
 - Sehr schwierig eine komplexe und rekursive Grammatik zu definieren
 - Lange Kompilationszeiten

- Spirit X3 hat angeblich eine gute Runtime Performance (keine verlässlichen Quellen, nur der Grundtenor von vielen Diskussionen)
- C++ mit PEGTL
 - Sehr ähnliche Probleme wie mit Spirit X3
- C++ mit ANTLR
 - Grammatik relativ einfach zu definieren, nah an der Backus-Naur-Form
 - Unterstützung von ANTLR in C++ ist nicht so ausgereift wie die von C#
 - Siehe die aufgeführten Nachteile von C++ in diesem Dokument
 - Implementierungs-Geschwindigkeit wichtiger als Runtime-Performance
- C# mit ANTLR
 - Grammatik relativ einfach zu definieren, nah an der Backus-Naur-Form
 - ANTLR hat die beste Interoperabilität mit C# von den unterstützten Sprachen welche ich beherrsche
 - Die Arbeit mit C# erlaubt einfache Nutzung einer Graphischen Oberfläche und hat auch sonst eine schnelle Implementierungsgeschwindigkeit

4.2 Details

Die Entstehung und die Komponenten eines Compilers ist meist in diese Schritte unterteilt:

- Design der Sprache (Siehe 3.2 Design der Sprache Myll, Seite 19)
- Lexikalische Analyse mittels Lexer
- Syntaktisch Analyse mittels Parser
- Semantische Analyse mittels Sema
- Ausgabe des Kompilats
- ???
- Profit :-P

Auch Myll hat diese Schritte, in diesem Fall der Transpilation ist die Ausgabe des Kompilats eine Erzeugung von C++ Quellcode.

4.3 Lexer / lexikalischer Scanner

Lexer werden verwendet um einen Eingabetext in logisch zusammenhängende Stücke zu zerteilen, diese Stücke werden Tokens genannt. Dieser Schritt, welcher *lexikalische Analyse* genannt wird, dient der Vorverarbeitung der Eingabe und wird von nahezu jedem Compiler durchgeführt. Leerraum wird im Normalfall in diesem Schritt verworfen (außer bei Sprachen in denen Leerraum Bedeutung hat, wie etwa Python). Oft werden Lexer durch ein Tool erzeugt, diese nehmen ein Vokabular und erzeugen daraus Programmcode. Alternativ zur Generierung können sie aber durchaus auch von Hand erstellt werden.

Ein Beispiel:

```
var int test = 99;
```

Diese Eingabe wird in folgende mit `|` getrennte Tokens aufgeteilt, Klammern zeigen optionalen Inhalt:

```
VAR | INT | ID( test ) | ASSIGN | INTEGER_LIT( 99 ) | SEMI
```

Der für Myll implementierte Lexer hat nur wenige Besonderheiten, zu seiner Erstellung wurde der Lexer-Generator von ANTLR 4 zu Hilfe genommen wurde, hierzu musste ein Vokabular definiert werden.

Die üblichen Spezialfälle wie *String* und *Character-Literale* sind so definiert das sie an ihrem einleitenden Zeichen anfangen (»"« und »'«). Von da an werden alle folgenden Zeichen und die meisten Escape-Sequenzen zum Literal gezählt, bis die Verarbeitung auf ein schließendes Zeichen stößt, welches identisch mit dem respektiven einleitenden Zeichen ist. Das Character-Literal muss es genau ein Zeichen enthalten.

Die Escape-Sequenzen für Oktale-, Hexadezimale- und Unicode-Zeichen werden bislang nicht unterstützt. Alternativ dazu ist aber die direkte Eingabe von UTF-8 Zeichen im Texteditor möglich.

Auch *Kommentare* sind in Myll möglich und erlauben die aus C++ bekannten Varianten, *bis zum ende der Zeile* via `// ...` und *über mehrere Zeilen hinweg* via `/* ... */`. Sie schreiben alle Eingaben in ihrem Geltungsbereich in einen speziellen Kanal, welcher abseits der zu verarbeitenden Code-Tokens liegt. Aktuell wird dieser Kanal noch nicht weiter genutzt, d.h. die Kommentare werden nicht in den erzeugten übernommen, es ist aber durchaus möglich dies zu einem späteren Zeitpunkt nachzurüsten.

Die Sprachen C sowie C++ leiden unter einem Entscheidungsproblem welches ich in Error: Reference source not found schon aufgeführt habe. Die Problematik ist, dass hier Identifier (Bezeichner) schon während der lexikalischen Analyse in *Typen Identifier* und *Variablen Identifier* klassifiziert werden müssen. Zur Lösung dieser Problematik wird ein sogenannter *Lexer-Hack* verwendet um die Tokens richtig zu klassifizieren. Darin wird dem Lexer vom Parser ein Rückwärtskanal in Form einer Symboltabelle bereitgestellt, anhand derer eingesehen werden kann ob ein Identifier nun einen Typen oder eine Variable darstellt.

(Siehe https://en.wikipedia.org/wiki/Lexer_hack)

Clang verzichtet auf diesen Hack und klassifiziert zunächst alle Identifier in eine gemeinsame Kategorie und entscheidet weit später, während der semantischen Analyse, die richtige Klassifikation.

Myll braucht diesen Hack ebenfalls nicht und unterscheidet Identifier auch nicht. Es ist durch die eindeutigere Syntax aber schon im Parser bekannt ob es sich um einen Typ oder eine Variable handelt.

Der für Myll implementierte Lexer betrachtet aber auch weiterhin die Tokens `>>` nicht gemeinsam, sondern als jeweils zwei separate `>` Tokens, so wie C++. Dies ist notwendig da sie je nach Auftrittsort entweder der logische bitweise Verschiebung nach rechts oder das schließen zweier Templates gemeint sind. Dies ist außerdem nun auch für den neuen Potenzierungs-Operator `**` nötig, da es in anderem Kontext auch zwei Dereferenzierungs-Aufrufe sein können.

```
vector< vector< int >> // Zwei separate Templates werden geschlossen
      84 >> 1 // Logischer rechts-shift
8 ** 2 // 8 hoch 2
**a // a wird zwei mal dereferenziert
```

Nach dem Schritt der lexikalischen Analyse ist bekannt ob der eingegebene Text Anomalien aufweist. Keine Anomalien heißt aber noch lange nicht das ein Text Sinn macht, also ein syntaktisch korrektes Programm zeigt.

```
if if int ()); // OK vom Lexer: Keine Anomalie
€ // Lexer meldet Fehler: € ist weder Keyword, Operator oder Identifier
" // Lexer meldet Fehler: Das Anführungszeichen wird nicht geschlossen
```

Die komplette Auflistung des Vokabulars des Lexers ist im Anhang ??? einzusehen.

4.4 Parser / syntaktische Analyse

Parser nehmen den vom Lexer erzeugten Token-Strom und erzeugen daraus entweder einen AST (Abstract Syntax Tree, einen Syntax Baum) oder einen Parse Tree (auch CST, Concrete Syntax Tree).

Mit dem linearen Strom an Tokens welcher der Lexer erzeugt hat, durchläuft der Parser in der syntaktischen Analyse meist einen Entscheidungsbaum, in jenem er versucht alle Tokens zu passenden Regeln zu zuordnen. Die definierten Regeln sind mittels *Und* und *Oder* Gliedern verkettet und der gesamte Regelsatz wird Grammatik genannt. *Und* Verknüpfungen deuten an das mehrere Tokens in Reihe übereinstimmen müssen, *Oder* Verknüpfungen bieten alternative Pfade an, trifft eine Regel nicht zu wird die nächste versucht. Jede Regel, die mit dem aktuell betrachteten Token übereinstimmt, wird in die Tiefe des Baums gefolgt. Sollte eine Regel in keinem Pfad zu einem Token passen, wird im Baum wieder nach oben gesprungen und von dort aus Alternativen durchlaufen. Dabei werden alle erfolgreich übereinstimmenden Regeln aufgezeichnet. Sollten bis zum Ende der syntaktischen Analyse alle Tokens verbraucht worden und der Regel-Baum wieder am Wurzelknoten sein, so bilden diese aufgezeichneten Knoten den AST oder CST des eingegebenen Textes und bestätigen das der eingegebene Text gültig ist. Wie Lexer, werden auch Parser oft von einem Tool erzeugt. Deren händische Erstellung ist ebenso möglich aber, je nach Komplexität der gewünschten Grammatik, ein sehr ambitioniertes Unterfangen.

Der für Myll implementierte Parser wurde mittels des ANTLR 4 Parser Generators erstellt. In jenem kann eine kontextfreie Grammatik in Erweiterter Backus-Naur-Form (EBNF) formuliert werden. ANTLR erzeugt seit Version 4 keinen AST mehr, sondern einen CST.

Anhand der folgenden stark reduzierten Grammatik werden die beispielhaft erzeugten Tokens aus dem Lexer Beispiel syntaktisch analysiert. Tokens werden hier komplett groß geschrieben, Regeln starten mit einem Kleinbuchstaben.

```

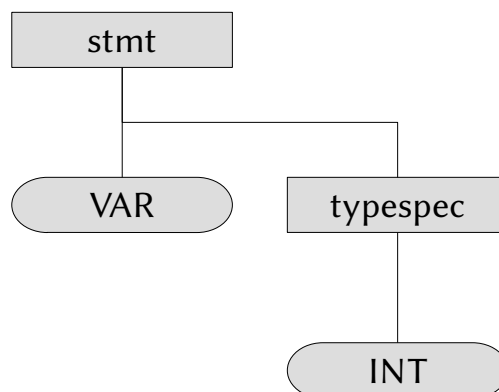
stmt      : ( RETURN expr SEMI
            | VAR typespec idVars SEMI );
typespec  : ( BOOL
            | FLOAT
            | INT );
idVars    : idVar (COMMA idVar)*;
idVar     : ID (ASSIGN expr)?;
expr      : ( expr PLUS expr
            | ID
            | INTEGER_LIT );

```

Der Token Strom des Lexer Beispiels

```
VAR | INT | ID( test ) | ASSIGN | INTEGER_LIT( 99 ) | SEMI
```

Der Einstiegspunkt in der Grammatik sei für dieses Beispiel die Regel `stmt` und das aktuelle Token ist `VAR`. Die Regel `stmt` wird betrachtet und enthält als erste Alternative `RETURN` (es ist eine Alternative weil die beiden Regeln mit einem *Oder* getrennt sind), dies stimmt nicht mit dem aktuellen Token überein also wird jetzt die Alternative betrachtet. Diese enthält `VAR` als erstes Element, was mit dem aktuellen Token übereinstimmt, das heißt der Token Strom springt zum nächsten Token. Zwischen `VAR` und `typespec` steht kein Operator, deswegen ist es eine implizite *Und* Verknüpfung. `typespec` ist eine weitere Regel und kein Token, deswegen wird sie zur weiteren Überprüfung betreten. In `typespec` wird jetzt geschaut ob ein Token mit `INT` übereinstimmt (natürlich der Reihe nach), und ja ist es: Strom einen weiter. Da `typespec` keine weiteren Regeln mehr enthält gilt es als erfolgreich erfüllt.

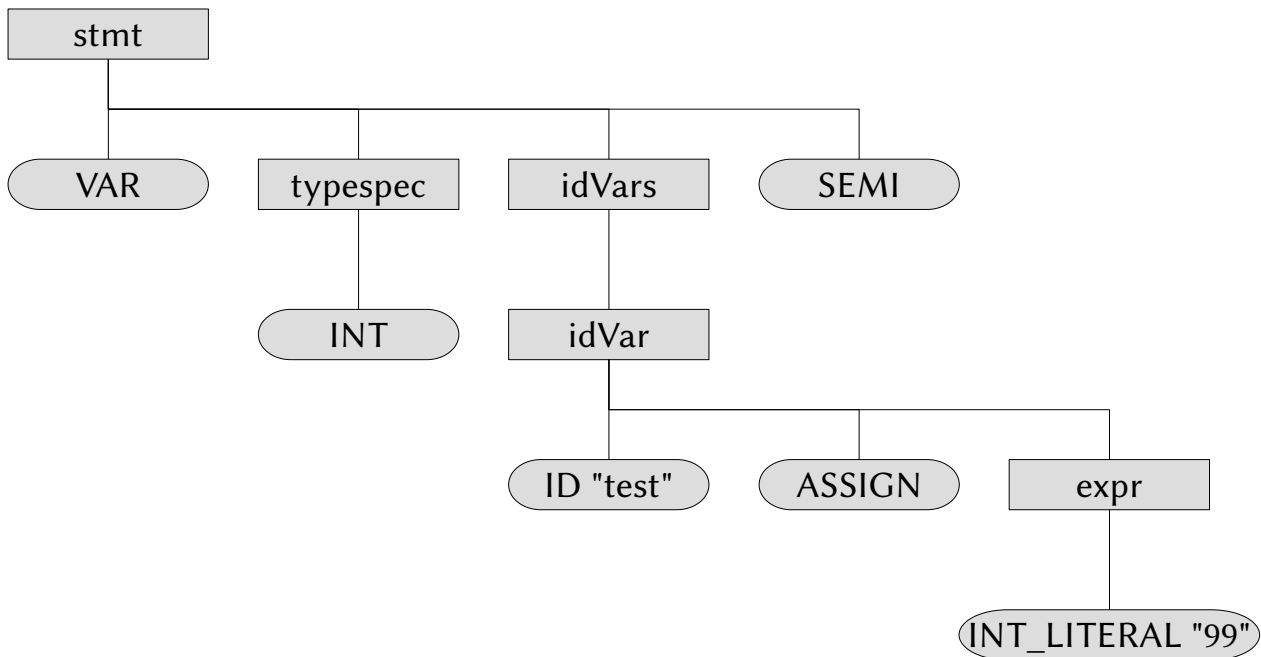
Illustration 4.1: Der Parse Tree / CST sieht aktuell so aus

Zurück im Ursprung in `stmt` steht `idVars` als nächste Regel an, diese wird betreten und es findet sich eine weitere Regel `idVar`, auch dort wird hineingesprungen. `ID` stimmt mit der `ID` aus dem Strom überein, der textuelle Inhalt des Token ist hier egal, die Regel ist erfüllt, Strom++. Dann kommt ein optionaler Anteil (zu erkennen am Fragezeichen als Suffix) welcher erfüllt sein kann, aber nicht muss. Es wird überprüft ob der nächste Token mit `ASSIGN` übereinstimmt, was er macht, Strom++. Der optionale Bereich ist aber noch nicht vorbei, d.h. die Regel `expr` wird betreten, welche wiederum Alternativen enthält. Die erste Regel enthält wieder `expr` als erste Regel. Aktuell sieht es anscheinend so aus als würde dies in einem ewigen Abstieg durch immer wieder der ersten (linken) Regel machen. Glücklicherweise hat ANTLR 4 genau für diese Problematik eine Sonderbehandlung, welche die direkte Rekursion der Regel ganz links erlaubt, welche bei vielen anderen Parsergeneratoren jetzt dazu geführt hätte das man seine Grammatik auf komplett andere Weise neu implementieren müsste. Siehe: Precedence climbing ???

(https://en.wikipedia.org/wiki/Operator-precedence_parser#Precedence_climbing_method)

Die Sonderbehandlung führte dazu das die erste Regel temporär übersprungen wurde und nun `INTEGER_LIT` mit `ID` verglichen wird, was nicht übereinstimmt. Die nächste Alternative ist `INTEGER_LIT`, ein Treffer also Strom++ und damit ist die optionale `(ASSIGN expr)?` Regel erfüllt, als auch `idVar` beendet. Jetzt steht eine optionale als auch wiederholbare Regel an (Erkennbar am Asterisk als Suffix), welche mit einem `COMMA` beginnen soll; hier könnten weitere Variablen vom gleichen Typ definiert werden. Diese Regel ist nicht erfüllt und `idVars` ist damit abgeschlossen. In `stmt` wird noch verlangt das noch ein `SEMI` folgt. Wäre dies nicht der Fall, wäre alle Arbeit verworfen worden und der Text als ungültig bewertet worden.

Illustration 4.2: Finaler Parse Tree / CST



Nach dem Schritt der syntaktischen Analyse ist bekannt ob der eingegebene Text grammatisch valide ist, was noch nicht heißt das dessen Code so auch ausgeführt werden kann.

Die vollständige Auflistung der Grammatik des Parsers ist im Anhang ??? einzusehen.

4.5 Sema / Semantische Analyse

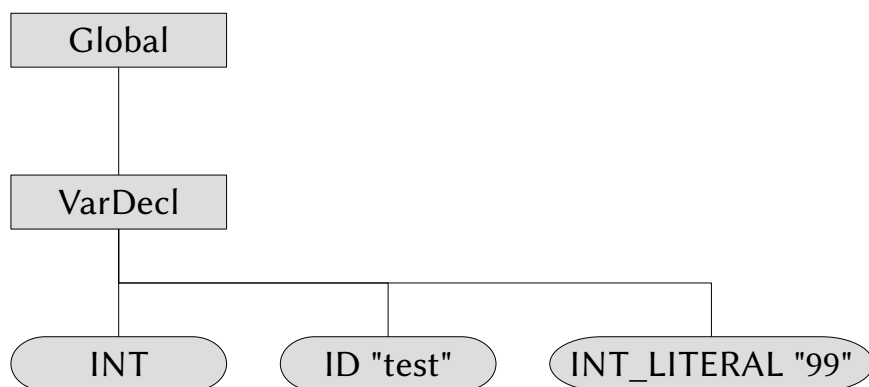
Die semantischen Analyse betrachtet die Bedeutung des Textes ... Allgemein???

Clang nennt die für die semantische Analyse genutzte Softwarekomponente Sema, dieser Namensgebung möchte ich mich hier anschließen.

Für die semantische Analyse gibt es keine Generatoren, zumindest nicht von ANTLR.

Da ANTLR anstelle eines AST, einen CST erzeugt hat, wird dieser Schritt hier nachgeholt. Der Unterschied eines AST zu einem CST besteht darin, das ein AST meist nur Blätter und Verzweigungen enthält, während der Baum eines CST viele unnötige Zwischenschritte beinhaltet. So sieht der AST des zuvor erzeugten Beispiels aus.

Illustration 4.3: Finaler AST



VarDecl hat zwei verpflichtende und einen optionalen Kindsknoten: Typ, Name und opt. Initialisierung. Der Informationsgehalt des AST entspricht dem des CST, die Informationsdichte des AST ist höher.

Bei der Frage wie der CST durchlaufen werden soll, bot ANTLR zwei Möglichkeiten: *Visitor* (Besucher) und *Listener* (Lauscher). Die Entscheidung fiel auf den *Visitor*, da dort der Durchlauf des Syntax Baums selbst gesteuert werden kann; man kann Visit() auf die Kindsknoten in einer selbst gewählten Reihenfolge aufrufen oder auch manche Teile gar nicht besuchen. Alternativ dazu hätte der *Listener* alle Knoten von selbst in vorgegebener Reihenfolge durchlaufen und einem lediglich die Möglichkeit gegeben dabei zuzuhören.

Die Semantische Analyse soll in Myll in drei Schritte unterteilt werden, der erste dieser Schritte findet schon beim Traversal des CST statt. Mehr???

Neu:

1. Durchlauf des CST, Erzeugung eines vorläufigen AST des kompletten Programms.
Die Namen von global sichtbaren Deklarationen werden in einer Datenstruktur erfasst:
 - Namespaces
 - Typen Deklarationen (struct, class, enum, union, alias)
 - globale-, statische- und instanz-Variablen
 - Funktionen und Methoden

Alle importierten Module müssen diesen Schritt durchlaufen haben, bevor der nächste Schritt beginnen darf, weil sonst Typen womöglich falsch aufgelöst werden.

2. Alle globalen Namen sind bekannt, alle Referenzen auf Globale können jetzt zugeordnet werden.
Diese Referenzen sind:
 - Using Namespace
 - Basisklassen
 - Typen der Variablen
 - Parameter und Rückgabetypen von Funktionen und Methoden

Die Zuordnung geschieht durch die Ersetzung der textuellen Repräsentation durch einen Zeiger auf das aufgelöste Element.

Aus myUsingDecl { namespace = "std" } wird myUsingDecl { namespace = &myStdDecl }

!!! Dieser Schritt ist nicht vollständig implementiert im aktuellen Prototypen

TODO: Fähigkeit diese Daten in einfach einzulesender Form abzuspeichern. Für partielle rekompilationen können diese eingelesen werden ohne die abhängigen Module neu zu kompilieren. Auch nicht-Myll Code kann über diesen Weg bereitgestellt werden analog zu TypeScript .d.ts Dateien.

3. Funktionen, Methoden und Expressions können nun aufgelöst werden. Da Funktionskörper keine globalen Deklarationen mehr verändern, kann dieser Schritt massiv parallel ausgeführt werden.

??? HIER WEITER 16. Mai 2020

Was sind Module??? So etwas wie *Includes* in C++, aber Module eskalieren keine Imports.

- Mod a importiert Mod b
- Mod b importiert Mod c
- Mod c ist in Mod a nicht direkt verfügbar

Dies ist Gegensätzlich zu Includes, welche sämtliche *sub-includes* mitbringen. Man könnte sagen Module sind nicht transitiv, Includes jedoch schon. Dies ist analog zu Modulen in C++20.

Selbst wenn ich Schritt 2 noch nicht vollständig mache, könnte ich dennoch wissen was Pointer sind und was nicht und ich könnte einen Zugriff auf Circle* c; via c.area() erlauben und in c->area() übersetzen. (Siehe Go Lang)

In C++ gibt es nichts was die Syntax `c.irdingwas` auf einen Raw-Pointer sinnvoll nutzt.

Was bedeutet 'fertig'??? erklären???

Beispiel

```
var int g_i; // (1) g_i
func myfunc() -> BC::Sub { // (2) myfunc
    var BC a;
    var BC::Sub b;
    a.c( b, g_i );
    return b;
}
class BC : SomeBase { // (3) BC
    class Sub {} // (4) BC::Sub
}
class SomeBase { // (5) SomeBase
    func c<T>( BC::Sub& b, T i ) {...} // (6) SomeBase::c<T>
}
```

In diesem Beispiel sind sechs globale Deklarationen enthalten

Schritt 1:

- Alle globalen Deklarationen werden hierarchisch durchlaufen
- Die globale Variable namens `g_i` wird mit dem eingebauten Typen `int` erfasst, diese Deklaration ist damit schon vollständig aufgelöst und wird als fertig markiert
- `myfunc` wird global als Funktion erfasst mit `BC::Sub` als Rückgabetyt, da dieser bisher unbekannt ist, wird er aktuell nur textuell erfasst und die Implementierung von `myfunc` wird vorerst ignoriert
- `BC` wird global als Klasse erfasst, die Basisklasse `SomeBase` ist noch unbekannt und wird deswegen auch nur textuell erfasst
- `Sub` wird unter `BC` als Klasse erfasst, ist vollständig aufgelöst und wird als fertig markiert
- `SomeBase` wird global als Klasse erfasst
- Die Funktion `c` wird unter `SomeBase` mit Parametertyp `BC::Sub` textuell erfasst, `T` bleibt offen
- Implementierung von `SomeBase::c` wird vorerst ignoriert

Schritt 2:

- Alle nicht fertigen globalen Deklarationen werden erneut hierarchisch durchlaufen
- Der Rückgabetyt von `myfunc` wird jetzt durch einen Verweis auf die in Schritt 1 erfasste `BC::Sub` Klasse ersetzt und da ihre Signatur vollständig aufgelöst ist wird sie als fertig markiert
- Die Basisklasse von `BC` wird durch einen Verweis auf die Klasse `SomeBase` ersetzt, sie wird als fertig markiert, da hier nichts direkt auf die (bisher nicht fertige) Basisklasse `SomeBase` verweist
- Die Parametertyp `BC::Sub` von `SomeBase::c` wird aufgelöst, `T` bleibt offen, wird als fertig markiert
- Rückkehr zur Klasse `SomeBase`, diese wird als fertig markiert da alle Kinder fertig sind
- Schritt 2 wird ein erneut durchlaufen sollte etwas nicht als fertig markiert sein
- Sollte die Anzahl der noch nicht fertigen Deklarationen nicht zwischen den erneuten Aufrufen schrumpfen, liegt ein fehlerhaftes Programm vor

Schritt 3:

- Alle nicht-Template Funktionskörper werden aufgelöst
- Lokaler Scope von `myfunc` wird erzeugt

- Variable *a* vom Typ *BC* wird in lokalem Scope erzeugt
- Variable *b* vom Typ *BC::Sub* wird in lokalem Scope erzeugt
- Variable *a* ruft die Methode *c* auf, in dieser wird der Template-Parameter *T* mit dem Typen *int* der globalen Variable *g_i* instanziiert und der Typ der Variable *b* ist kompatibel mit dem ersten Parameter
- Variable *b* wird zurückgegeben, dessen Typ ist kompatibel mit dem Rückgabetyt der Funktion
- Der lokale Scope wird geschlossen
- Templatefunktionen wie *SomeBase::c* werden (wie in C++) nur im Falle der Nutzung aufgelöst. Im Falle ihrer Nutzung sind sie natürlich vollständig typisiert und werden wie normale Funktionen erzeugt.

4.6 Details zur Implementierung

Drei verschiedene Arten von Basistypen machen die Knoten meines AST aus:

Decl - Deklarationen (welche korrekterweise Globale- & Instanz-Deklarationen hätte heißen müssen)

Stmt - Statements (welche korrekterweise Lokale-Deklarationen und Statements hätte heißen müssen)

Expr - Expressions

Die allgemeine Verschachtelung dieser folgt den Regeln das Decl alle Knotentypen enthalten können, Stmt kann nur Stmt und Expr enthalten und Expr kann nur Expr enthalten. Eine einzelne Ausnahme für Stmt welches Decl enthalten kann mag gemacht werden, sollten lokal definierte Strukturen / Funktionen zum Funktionsumfang hinzugefügt werden, was aktuell nicht möglich ist.

4.6.1 Deklarationen

Wie schon erwähnt wird in Myll eine Deklaration nicht von der Definition getrennt, das bedeutet das überall wo Deklaration geschrieben ist, immer auch die Definition mit gemeint ist.

Diese Klassen machen alle Deklarationen aus:

- Decl – Basisklasse aller Arten von Deklarationen, *abstrakt*
 - Hierarchical – Alle Decls welche weitere Decls enthalten können, *abstrakt*
 - Structural – Allgemeine strukturelle Deklarationen: `struct`, `class` und `union`
 - Enumeration – Enum Deklaration
 - Namespace – Namespace Deklaration
 - GlobalNamespace – Ausgangspunkt eines jeden Moduls
 - Var – Globale Variablen, Globale Konstanten, Klassen-Variablen, automatische Accessoren
 - Using – Deklaration von Typ-Aliasen und Namespaces
 - Func – Deklaration von Funktionen, Methoden, Operatoren und manuelle Accessoren
 - ConDestructor – Deklarationen von Kon- und Destruktoren
 - EnumEntry – Eintrag in einem Enum
 - MultiDecl – Container welcher mehrere Decl in einer einzelnen Decl speichert

Der `GlobalNamespace` ist der Ausgangspunkt jedes Moduls und somit jedes AST, er entspricht damit quasi dem was in C++ Translation Unit genannt wird. `GlobalNamespace` enthält im Vergleich zu `Namespace` noch den aktuellen Modulnamen und die Imports.

`MultiDecl` war nötig um aus einem ANTLR `visit()` mehr als ein Element zurückzugeben. Es gibt keine Notwendigkeit von Myll das diese Klasse existiert.

`EnumEntry` ist nur in Enumeration gültig, `ConDestructor` nur innerhalb von Structural (class und struct), `Namespace` darf nur innerhalb von `Namespace` und damit auch `GlobalNamespace` erscheinen.

4.6.2 Statements

Statements können in Funktionskörpern und als Teil von anderen Statements auftreten.

- Stmt – Alle Arten von Statements
 - VarStmt – Deklaration lokaler Variablen und lokaler Konstanten
 - UsingStmt - Deklaration lokaler Typ-Aliasen und Namespaces
 - IfStmt – Kontrollstruktur zur Verzweigung
 - SwitchStmt - Kontrollstruktur zur Mehrfachverzweigung
 - CaseStmt – Sprungziel einer Mehrfachverzweigung
 - LoopStmt – Basisklasse aller Schleifen, ist selbst eine endlose Schleife
 - ForStmt – Übliche drei Komponenten For-Schleife: Initialisierung, Bedingung, Fortschritt
 - WhileStmt – Kopfgesteuerte Schleife
 - DoWhileStmt – Fußgesteuerte Schleife
 - TimesStmt – Zählschleife
 - ContinueStmt – Bricht aktuelle Iteration ab und fährt mit der nächsten fort
 - BreakStmt – Abbruchs Statement für Schleifen und Mehrfachverzweigungen
 - FallStmt – Designiert ein gewünschten Fallthrough bei Mehrfachverzweigungen
 - ReturnStmt – Rücksprung aus einer Funktion
 - TryCatchStmt – Ausnahme Behandlung
 - ThrowStmt – Ausnahme Auslösung
 - MultiAssign – Ein oder mehrere Zuweisungen, z.B. `a = b = true;`
 - AggrAssign – Eine aggregierende Zuweisung, z.B. `a += 1.41f;`
 - Block – Ein neuer Scope welcher mehrere Statements enthalten kann
 - MultiStmt – Mehrere Statements ohne neuen Scope
 - EmptyStmt – Leeres Statement
 - ExprStmt – Ein Statement welches nur Expressions enthält

`MultiStmt` war nötig um aus einem ANTLR `visit()` mehr als ein Element zurückzugeben. Es gibt keine Notwendigkeit von Myll das diese Klasse existiert. `EmptyStmt` existiert als Lückenfüller für Stellen an denen ein Statement optional wäre.

4.6.3 Expressions

Expressions bestehen aus der größten Menge an unterschiedlichen Arten im gesamten AST. Um diese extreme Diversität nicht ausufern zu lassen, bleiben einige von ihnen ohne spezifische Subklasse, sie sind von den Typen UnOp & BinOp und ihr effektiver Typ ist in einem Enum codiert. Dieses Enum erfüllt auch einen wichtigen Zweck um die Änderung in der Operator Precedence bei der Erzeugung der Ausgabe zu steuern.

Die Klassen die von Expr ableiten sind:

- Expr – Alle Arten von Expressions, *abstrakt*
 - UnOp – Unäre Operationen welche eine weitere Expression enthalten, z.B. `-e`, `e++`, `sizeof(e)`
 - CastExpr – Typenumwandlung der enthaltenen Expression, z.B. `(? MyType) e`, `(move) e`
 - FuncCallExpr – Aufruf der Expression als Funktion/Methode, z.B. `e(arg, arg)`
 - BinOp – Binäre Operation welche zwei weitere Expressions enthalten, z.B. `e * e`, `e.e`, `e == e`
 - TernOp – Ternäre Operation welche drei weitere Expressions enthalten, z.B. `e ? e : e`
 - Literal – Jegliche Literale / Werte, z.B. `42`, `true`, `"test"`, `0xACAB`, `this`
 - IdExpr – Identifier mit optionalem Template Bestandteil, z.B. `myVar`, `pow<float>`
 - ScopedExpr – Verschachtelte IdExpr, z.B. `MyClass::staticField`, `std::is_pod<Typ>`
 - NewExpr – Speicheranforderungs Operation, z.B. `new Balloon<0xFF0000>[99]`

Idsf

Attribute

4.6.4 Was lief nicht so gut (Überschrift verbessern???)

Die schlechteste Entscheidung die ich bei dieser Modellierung getroffen habe, ist die wenigen Überschneidungen der Deklarationen und Statements vom gleichen Code / den gleichen Klassen behandeln zu lassen. Dies betraf insbesondere die Deklaration der Lokalen-, Globalen- und Instanz-Variablen.

Dies kam Wahrscheinlich zustande weil Decl von Stmt abgeleitet ist. Decl muss zwar alles können was Stmt auch kann, erfüllt aber nicht die Anforderung *Decl ist ein Stmt*, zumindest nicht in Mylls Definition.

Problematisch waren einige convenience Schreibweisen, welche etwa mit einer Decl mehrere Decl erzeugten. Dies bot Probleme mit der Funktionsweise von ANTLR, welches als Rückgabe nur eine einzelne Decl erlaubte. Umschifft wurde dieses Problem durch die Einführung einer Multi-Decl Klasse welche lediglich ein Container für mehrere Decl ist.

4.7 Generator

Problem: Ich brauche eine spezialisierte String Klasse / Datenstruktur welche die effiziente Erzeugung von Output unterstützt.

Eckpunkte:

- Erzeugung von Iteratoren auf Positionen im Text, welche sich bei Einfügeoperationen mit verschieben und Reallokationen mitmachen
- Einfügen von Text an jeglicher Stelle der genannten Iteratoren, ohne das für jeden Aufruf alle Folgezeichen verschoben werden müssen
- Unterstützung von sehr großen Strings und mehrerer dessen parallel

Was ich nicht brauche:

- Thread-Sicherheit

4.7.1 Was lief gut

Parallelisierung lies sich für ein Großteil des Compilers relativ einfach umsetzen. Ein Teil des Compilers benutzt statische Member weshalb hier aktuell nicht parallelisiert werden konnte.

5 Auswertung & Zusammenfassung

5.1 Bewertung der Grundsatzeinhaltung

Wiederholung der Grundsätze:

5.1.1 Erwarte keine Wiederholung vom Benutzer

- Deduplizierung durch Reduktion auf eine Datei
- Keine Notwendigkeit von Prototypen

5.1.2 Außergewöhnliches Verhalten muss Explizit sein

- Konstruktoren Implicit
- Switch Fall
- Pointer auf Arrays / Array Decay

5.1.3 Das was man häufig will kann Implizit sein

- Kein Shadowing
- Switch Break
- Public Ableitung
- Funktionen sind standardmäßig noexcept
- Bitweise Operationen vor Gleichheit

5.1.4 Breche nicht mit der Semantik von C++

- Größte Sorgfalt wurde darauf gelegt die Semantik beizubehalten

5.1.5 Breche mit C sofern es einen Nutzen bringt

Jeder Bruch mit der Syntax von C ist dokumentiert und dient einem Nutzen.

- Keine undurchsichtige Funktionspointer und Array Syntax
- Keine unscoped Enumerationen
- Kein archaischer Präprozessor
- Keine Typen mit Leerzeichen getrennten Namen z.B. unsigned long long int
- Kein NULL und myPtr = 0

5.1.6 Entwickle die Syntax so weiter das sie Eindeutig bleibt

Viele syntaktische Makel welche in C & C++ auftraten wurden beachtet und vermieden

- Keywords für Variablen und Funktionen
- Most Vexing Parse (Erklären und Quelle, `vector<array<100>>4>>`)

5.1.7 Sei auch einmalig Nützlich

Dies kann etwa durch die Erzeugung von vom Menschen lesbarem C++ Code passieren.

5.1.8 Spare nicht mit neuen Keywords

Es wurden einige Keywords eingeführt um Ambiguitäten zu unterbinden, den Code Lesbarer zu gestalten und neue Konstrukte mit angenehmer Syntax einzuführen

5.2 Zieleinhaltung

Das Entscheidungsproblem unter dem C++ leidet, ob etwas eine Variablendefinition, eine Funktionsdeklaration oder eine mathematische Expression ist lässt sich durch die Einführung der *var* und *func* Keywords lösen. Es macht es auch fürs Menschengenie leichter entscheidbar was man gerade liest. Siehe 3.3.2.1

In Voraussicht wurde der Right-Shift-Operator in der Grammatik als zwei separate `>` definiert, welches die Probleme beim schließen zweier Templates löste. Ein Problem mit Templateargumenten trat dennoch auf, das etwa diese Expression hier falsch *verstanden* wurde:

```
static_cast< Type >( lhs ) | static_cast< Type >( rhs );  
static_cast<(Type >( lhs ) | static_cast< Type>( rhs ); // eine große Expression
```

Der Ursprung war aber ein komplett anderer als die alte Template Problematik, hier war nämlich definiert das ein Templateargument eine beliebige Expression sein konnte und diese Verhält sich *greedy* solange noch ein schließendes `>` vor dem nächsten `;` auftaucht.

Myll bietet leichtere Lesbarkeit dank weniger Einrückung durch die Fähigkeit verschachtelte Namespaces in einem einzelnen Block zu schreiben oder den Block gar ganz wegzulassen und den Rest der Datei in diesem Namespace zu haben. Es erwartet weniger Boilerplate dadurch das nicht bei jeder Definition `template <typename>, ClassName::` und eine erneute Aufzählung der Funktions- und Methodensignaturen gefordert wird. Einige Teile erfordern zwar mehr zu schreiben als bisher, allen Vorweg `func` und `var`, sie bieten aber auch leichtere Lesbarkeit weil sie einem die kognitive Last nehmen den kompletten Inhalt einer Codezeile zu lesen um zu wissen was dort generell passiert.

Myll kann sich leicht fortentwickeln da die Attributsyntax nicht mit Keywords und Identifiern kollidiert,

Die Problematik eine so komplexe Sprache wie C++ zu lesen, zu verstehen und dann wieder auszugeben ist ein hartes aber machbares Unterfangen. Ohne mir schon vor dieser Arbeit viel über C++ und allgemeinem Sprachdesign angeeignet zu haben, wäre es aber nicht in dem Zeitrahmen dieser Arbeit machbar gewesen.

Jemand anderes der ausgehend von dieser Arbeit einen Transpiler nach C++ aber auch C#, Java, Rust oder ähnlichem schreiben will, wird keine großartigen Probleme haben, solange die Grundstruktur der Sprache übereinstimmt.

5.3 Das sieht ja gar nicht mehr wie C++ aus!

Das stimmt schon... Aber sieht denn C++11 noch so aus wie C++98?

Nein, siehe folgendes Beispiel:

VALIDIEREN!!!

```
// C++98
int * pi = new int(42);           // Hier der Standard, verpönt in der Zukunft
delete pi;                       // Freigeben nicht vergessen!
typedef a b;                     // Typalias
void genVec(vector<int> outVec) {...} // Gute Performance nur mit pass-by-reference

// C++11
unique_ptr<int> pi = make_unique<int>(42); // So soll es in C++11 aussehen, delete unnötig
using b = a;                             // Moderner Typalias
vector<int> genVec() {...}                // OK in C++11, perfekt in C++17

Mehr BSPL???
perfect forwarding, move semantics
oder einfach einen guten Artikel verlinken?
```

Und wird C++20 noch so aussehen wie C++11?

Auch Nein: Der Standard ist schon abgesegnet und einiges von diesem neuen Standard hat sich im Vergleich zum letzten Großen neun Jahre vorher verändert. Es gibt schon Compiler welche viele Funktion des neuen Standards unterstützen.

Das Feature constexpr hat schon viel TMP unnötig gemacht und wird dies Dank der Einführung von constexpr fortsetzen. Man kann damit beispielsweise Funktionen schreiben welche zur Compiletime ausgeführt werden, die selben aber auch zur Runtime verwenden. Kein

Concepts werden die am schwersten zu durchschauenden Template-Parametrierungs-Fehler durch Sinnvolle Fehlermeldungen ersetzen und einige schwer zu schreibende SFINAE TMP Konstruktionen unnötig machen.

Der Spaceship Operator <=> wird einem die Arbeit abnehmen die meisten Vergleichsoperatoren zu implementieren. Er verhält sich relativ analog zu *strcmp*.

Es ist auch gut das man vom einen C++ Standard zum anderen eine Veränderung sehen kann, sonst hätte sich ja auch nicht viel verbessern können.

Genau das versuche ich auch zu Bezwecken, Myll soll *Vertraut* aber dennoch *Anders* sein. Vielleicht in eine andere Himmelsrichtung Anders als die kommenden C++ Standards, aber sonst wäre Myll ja auch sinnlos.

Known issues

(Partial) Template Specialization might not work, thus TMP might not work as well

Keine Lambdas ATM

Es wird Code geben der von Myll nach C++ kompiliert, dann aber beim kompilieren des C++ Codes scheitert. Aufgrund von unvorhergesehener Namenskollision oder ähnlichem. Wenn man Myll in dieser Form scheitern lassen will, dann schafft man dies auch.

Letter to:

Bjarne Stroustrup,

Tell what I do and why I ask.

I know that you do not like C# because of its proprietary nature. Thus I am only interested in your opinion on just C#'s Syntax. Especially in the light of your sentence: Cleaner Language...

I personally think that C#'s syntax for defining template classes and functions is almost always as expressive as C++'s and way easier to use. Can you relate to my gut feeling?

A. Heijlsberg,

Tell what I do and why I ask.

Do I even have Questions?
 A. Alexandrescu,
 I recently read an Answer on Quora concerning the DLanguage
 Herb Sutter
 Parameter passing, sink, passing shared, unique and raw pointers
 Per Vognsen,
 Jonathan Blow:
 NO NAGGING stuff...

Read the surroundings of the quotes.

5.4

Das Entscheidungsproblem unter dem C++ leidet, ob etwas eine Variablendefinition, eine Funktionsdeklaration oder eine mathematische Expression ist lässt sich durch die Einführung der *var* und *func* Keywords lösen. Es macht es auch fürs Menschenauge leichter entscheidbar was man gerade liest. Siehe 3.3.2.1

In Voraussicht wurde der Right-Shift-Operator in der Grammatik als zwei separate `>` definiert, welches die Probleme beim schließen zweier Templates löste. Ein Problem mit Templateargumenten trat dennoch auf, das etwa diese Expression hier falsch *verstanden* wurde:

```
static_cast< Type >( lhs ) | static_cast< Type >( rhs );  
static_cast<(Type >( lhs ) | static_cast< Type >( rhs ); // eine große Expression
```

Der Ursprung war aber ein komplett anderer als die alte Template Problematik, hier war nämlich definiert das ein Templateargument eine beliebige Expression sein konnte und diese Verhält sich *greedy* solange noch ein schließendes `>` vor dem nächsten `;` auftaucht.

Die Problematik eine so komplexe Sprache wie C++ zu lesen, zu verstehen und dann wieder auszugeben ist ein hartes aber machbares Unterfangen. Ohne mir schon vor dieser Arbeit viel über C++ und allgemeinem Sprachdesign angeeignet zu haben, wäre es aber nicht in dem Zeitrahmen dieser Arbeit machbar gewesen.

Jemand anderes der ausgehend von dieser Arbeit einen Transpiler nach C++ aber auch C#, Java, Rust oder ähnlichem schreiben will, wird keine großartigen Probleme haben, solange die Grundstruktur der Sprache übereinstimmt.

5.5 Was konnte nicht behandelt werden

Einige kleinere Punkte welche nicht umgesetzt wurden, sind schon in Kapitel 3.3 beschrieben.

5.5.1 Alles was der C++ Semantik widerspricht

Das Grundkonzept, welches ein hohes Maß an Machbarkeit dieses Projektes ausgemacht hat, ist dass die Semantik von C++ bewahrt wird. Wäre ein stärkerer Bruch mit C++ beabsichtigt worden, wäre die Implementation dieses Compilers wesentlich schwerer gefallen.

Ein Bruch mit der Semantik von C++ hätte sicher an einigen Stellen positive Auswirkungen haben können, aber das Grundkonzept wollte auch C++ Programmierer mit geringem kognitivem Umstellungsaufwand abholen. Eine Verletzung dieser Grundlage würde sich negativ auf dieses Ziel auswirken.

5.5.2 Adresse-Von- und Indirektions-Operatoren

Ein Problem welches mich ursprünglich vom *Verstehen* von C++ abgehalten hatte, war dass um aus einem Objekt einen Pointer zu machen, muss man `&` verwenden und nicht `*`. Umgekehrt, also wenn man schon einen Pointer (`*`) hat, muss man noch einen weiteren *Pointer* dazuschreiben damit der Pointer weggeht; Naivität beabsichtigt.

Eine logische Entwicklung wäre die Operationen `&` und `*` zu vertauschen um eine naiv logischere Situation zu erzeugen. Es würde allerdings auch harsch nicht nur mit der Homogenität zu C++ brechen, denn auch D, Rust, Go, C# (ja *auch* C# kann Pointer), etc verwenden die gleiche Syntax.

5.6 Performance Benchmark

Ruby Script um die Benchmark-Datei zu erzeugen:

```
f=File.read("stack_proto.myll");16.times{|n|fo=File.open "stack_big_"+n.to_s+".myll","w";fo.write "module
stack_big_"+n.modulo(4).to_s+";\n";500.times{|i|fo.write "namespace Rethmann::Uebung_"+i.to_s+";\n";fo.write f}; fo.close}
```

Parallelisierbarkeit

==== [#IN](#) ====

Genereller Scope von Myll:

- Einführung geplanter C++17/20 features ahead of time.

- Einführung nicht von ISO C++ geplanter Features.

- Änderung alter C Syntax, Modernisierung auf aktuelles C++ Verhalten. (Typen, Deklarationen)

- Einführung von Funktionalität welche nur im Transpiler existiert, für warnings oder errors verwendet wird, im Endeffekt aber keinen C++ Code erzeugt.

Vielfache Nutzung von static für unterschiedlichste Zwecke

- Beschränkung der Nutzung und ersetzen durch Alternativen, Bestenfalls als Attribut, kollidiert nicht mit der Sprache.

Different operator precedence table

- Boolsche Operatoren sind schlecht einsortiert: `a&1==0` ist `a&(1==0)` sollte `(a&1)==0` sein.

- Idee von /PV.

- Neue Operatoren, von C# und selbst Ausgedachte.

All variables initialized by default, uninitialized must be specified.

- Idee von /JB

Default relational ops and compound assignment

- Overrideable

Header and Implementation in a single file

Assignment is a Statement where left hand side Type communicates with the right hand side

- Wirkt effektiv gegen `if(a=1)`, macht anderen code auch wesentlich schwerer obfuszierbar.

Enums can be linear or quadratic auto-indexed (flags enum)

Enums have reflection back to their string representation, oder andersherum

- Leichter zu debuggen oder configs einzulesen.

Easier safe cast Expressions

Reuse the easy cast syntax from C, but this is only `static_cast`, others equally simple

Easier use of Smartpointers

`int @! a; für unique_ptr<int> a;`

Easier Template syntax

C# zeigt das es auch anders geht.

Concepts Lite Lite

Make Concepts available for all the easy cases

Spaceship Operator

UFCS

Nur in einer Richtung: aus `foo(a)` wird das hier nutzbar `a.foo()`

Accessors

Getter/Setter incl. dem syntaktischen zucker welcher sie ohne `()` aufrufbar macht.

Implicit `ret` var

Functions/Procedures liefern automatisch eine Variable die `ret` heißt und dem Rückgabetyt entspricht.

Enum methods

6 Ausblick

Everything is constexpr, when it can be, you don't need to write it.

6.1 Benamte Parameter

NICE TO HAVE, DON'T WASTE TIME, DUDE!

```
func spawn_entity(bool is_visible) {...}
func spawn_entity(bool put_a_hat_on) {...}

spawn_entity(is_visible: true);
```

Diese beiden Funktionen würden in C++ durch ihre identische Signatur nicht gemeinsam existieren können. In Myll ist ein Aufruf mit `is_visible: true` jedoch eindeutig der ersten Variante zuzuordnen. Ein Aufruf dieser mit benannten (benannten) Parametern überladenen Funktion ohne die Nennung der Argumentnamen ist allerdings nicht mehr möglich, außer der ein Attribut markiert die primär/namenlos zu verwendende Funktion oder ein Invertiertes markiert die namenlos Untauglichen.

```
[primary/nameless] func spawn_entity(bool initially_visible) {...}
// ODER
[explicit/onlynamed]      func spawn_entity(bool put_a_hat_on) {...}

spawn_entity(true);        // ruft die primäre Funktion auf
spawn_entity(is_visible: true); // trotzdem schöner zu lesen, weil man sieht was das true
                             // bedeutet
```

Benannt == Benamt?

6.2 Intelligenterer Parameter Übername

Der Programmierer muss im allgemeinen mehr Gehirnschmalz in die Gestaltung der Übernahmeparameter stecken als man Naïv annimmt.

IDE: Jeder Typ spezifiziert zu Compiletime welche Kosten ein Copy, ein Move oder ein Referenzieren hat. Auch für Literals des Typen

Als Beispiel dient eine Funktion welche nur wenige Parameter lesen will:

```
func(  int          a,
      vector<int>    b,
      shared_ptr<image> c,
      unique_ptr<database> d
) -> void { ... }
```

An welchen Stellen treten hier Probleme auf und welche Möglichkeiten bieten sich diese zu unterbinden?

- Dieser Parameter ist unproblematisch. Hier ließe sich lediglich noch ein *const* hinzufügen, was an der Korrektheit nach Außen nichts ändert, es wirkt sich lediglich als nur-lesen Einschränkung nach innen aus.

- b. Dieser Parameter hat in dieser Form ein potentiell performance Problem, da der Vektor mit seinem kompletten Inhalt kopiert wird. Mir fällt nur ein guter Grund ein dies so zu schreiben und zwar wenn man den Vektor innerhalb der Funktion ändern will, sich diese Änderung aber nicht nach aussen hin auswirken soll.
- c. Höchstwahrscheinlich passiert hier das Richtige, es ist im Falle des nur-Lesens lediglich ein überflüssiges Inkrementieren und Dekrementieren des usage-counters möglich.
- d. Dies ist sehr wahrscheinlich ein Fehler, es sei denn man wollte wirklich dass die Funktion die Datenbank konsumiert, sprich das übergebene Argument invalidiert.

Bauen wir die Parameter so um das sie wahrscheinlich das Richtige tun.

```
func(  const int ODER int ODER int& ODER int*                a,
      const vector<int>& ODER vector<int>& ODER const vector<int>* ODER vector<int>*  b,
      const shared_ptr<image> ODER const shared_ptr<image>& ODER const shared_ptr<image>*
      shared_ptr<image> ODER shared_ptr<image>& ODER shared_ptr<image>*            c,
      unique_ptr<database>& ODER unique_ptr<database>* ODER database*              d
) -> void { ... }
```

Dieses Feature lässt sich möglicherweise mit TMP lösen, wird dann aber wahrscheinlich eine verbose Syntax haben.

Nein, denn es ließe sich mit viel SFINAE Magic maximal die Signatur der Funktion ändern, die Position des Aufrufs bleibt unangetastet... außer es soll zusätzlich auch um jedes Argument aller Call Sites gewrapt werden, weil wenn z.B. in der decl aus `T -> T*` wird muss an der call site aus `arg -> &arg` werden.

```
template <typename T>
void foo(typename look_t<T>::type t) {...}
...
foo(look(argument));
```

6.3 Ranges

Läuft das so?

```
int sum = container | filter | sort | reduce;
```

6.4 Rest

==== [#MAYBE](#) ====

- ‘maybe_const’, is an Aspect of a Function which will make it implement a const and a non-const version of itself. It is a
- Named function arguments, makes refactoring easier and calls like this easier to read: `DoSomething(false) -> DoSomething(checkBefore:false)`.
For bools even this syntactic sugar might be possible: `DoSomething(!checkBefore)`, `DoSomething(NOTcheckBefore)`, `DoSomething(checkBefore)`, `DoSomething(DOcheckBefore)`
- Glue Classes (with UFCS?)
 - Classes that tie two or more classes together and only exist when all classes are loaded
- Multithreaded compiler
- Be configurable for the user e.g. “@!” can be `std::unique_ptr` or a custom pointer class
- Enable aspect oriented programming through Attributes
- Warnings for many holes in Structs due to alignment/bad ordering
- Reimagined Parameter definitions (lend, sink)
 - Copy:
 - Look:

- Edi
- UTF8 in strings, 32bit Codepoint for single chars

==== [#OUT](#) ====

Use ifs everywhere as alternative to #ifdefs

Dynamic size for integer types: i23 for a 23 bit integer

Do not piggyback on C++ anymore

Nullable integer types which sacrifice a bit or less for being nullable. They can also be symmetric in positive and negative range. i7 could reach from -127 to +127 or -63 to +63 or anything in between.

Replace const and volatile with symbols (#,~,@,???).

Overloadable nearly-equals operator (~==) for floating point types, with as many special cases as you like to *pay* for.

7 Appendix

7.1 Literaturverzeichnis

[SPECS] Ben Werther & Damian Conway, *A Modest Proposal: SPECS*:

<http://users.monash.edu/~damian/papers/PDF/ModestProposal.pdf>

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.2481&rep=rep1&type=pdf>

[KR-CPL] Kerninghan & Ritchie, *The C Programming Language*:

[DOD] Mike Acton, *Data-Oriented Design and C++*:

<https://github.com/CppCon/CppCon2014/blob/master/Presentations/Data-Oriented%20Design%20and%20C%2B%2B/Data-Oriented%20Design%20and%20C%2B%2B%20-%20Mike%20Acton%20-%20CppCon%202014.pptx>

<https://www.youtube.com/watch?v=rX0ltVEVjHc>

[JAI] Jonathan Blow, *Ideas about a new programming language for games*:

<https://www.youtube.com/watch?v=TH9VCN6UkyQ&list=PLmV5I2fxaiCKfxMBrNsU1kgKJXD3PkyxO>

[ION] Per Vognsen, *Bitwise Ion*:

https://github.com/pervognsen/bitwise/blob/master/notes/ion_motivation.md

[https://www.youtube.com/watch?v=T6TyvsKo_KI&list=PLU94OURih-](https://www.youtube.com/watch?v=T6TyvsKo_KI&list=PLU94OURih-CiP4WxKSMt3UcwMSDM3aTtX&index=3&t=3100s)

[CiP4WxKSMt3UcwMSDM3aTtX&index=3&t=3100s](https://www.youtube.com/watch?v=T6TyvsKo_KI&list=PLU94OURih-CiP4WxKSMt3UcwMSDM3aTtX&index=3&t=3100s)

[SHADOW] Malte Skarupke, *Ideas for a Programming Language Part 3: No Shadow Worlds*:

<https://probablydance.com/2015/02/16/ideas-for-a-programming-language-part-3-no-shadow-worlds/>

[EPOCHS] Vittorio Romeo, *Epochs: a backward-compatible language evolution mechanism*:

https://vittorioromeo.info/index/blog/fixing_cpp_with_epochs.html

<http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2019/p1881r0.html>

[JM-EA] Jonathan Müller, *Write explicit constructors - but what about assignment?*

<https://foonathan.net/2017/10/explicit-assignment>

[JM-SMF] Jonathan Müller, *Tutorial: When to Write Which Special Member*:

<https://foonathan.net/2019/02/special-member-functions>

[HS-SPP] Herb Sutter, *GotW #91 Solution: Smart Pointer Parameters*:

<https://herbsutter.com/2013/06/05/gotw-91-solution-smart-pointer-parameters>

[LEXHACK] https://en.wikipedia.org/wiki/Lexer_hack

[ANTLR4] <https://www.antlr.org/papers/allstar-techreport.pdf>

[TR-R5Y] The Register, *Rust marks five years since its 1.0 release: The long and winding road actually works*:

https://www.theregister.co.uk/2020/05/15/rust_marks_five_years_since/

MAYBE TOO MUCH:

[ARENA] John Lakos, *Local ("Arena") Memory Allocators*:

<https://www.youtube.com/watch?v=CVbnjydW5M0>

7.2 Tabellenverzeichnis

Tabelle 3:1: Operator Precedence Table.....	35
---	----

7.3 Abbildungsverzeichnis

Illustration 4:1: Der Parse Tree / CST sieht aktuell so aus.....	49
Illustration 4:2: Finaler Parse Tree / CST.....	50
Illustration 4:3: Finaler AST.....	50

7.4 Listings

7.4.1 Vokabular des Lexers

```
lexer grammar MyllLexer;

channels { NEWLINES, COMMENTS }

COMMENT      :      (      '#!' ~('\r'|\n')*           // ignore shebang for now
                    |      '//' ~('\r'|\n')*
                    |      '/*' .*? '*/'
                    ) -> channel(COMMENTS);

STRING_LIT   :      '"' (STR_ESC | ~('\\" | '"' | '\r' | '\n'))* '"';
CHAR_LIT     :      '\'' (CH_ESC | ~('\\" | '\'' | '\r' | '\n')) '\'';
fragment STR_ESC:      '\\\' | '\\\' | '\'' | 'a' | 'b' | 'f' | 'n' | 't' | 'r' | 'v');
fragment CH_ESC:      '\\\' | '\\\' | '\'' | 'a' | 'b' | 'f' | 'n' | 't' | 'r' | 'v');

MOVE         :      '(move)';
ARROW_STAR   :      '-> *';
POINT_STAR   :      '. *';
PTR_TO_ARY   :      '[' *';           // [*] could be a dynamic array
COMPARE      :      '<=>';
TRP_POINT    :      '...';
DBL_POINT    :      '..';
DBL_AMP      :      '&&';
DBL_QM       :      '??';
QM_COLON     :      '?:';
//DBL_STAR   :      '**'; // this is only supported by 2x STAR because of: var int** a
//              // which is a double pointer, not a pow

DBL_PLUS     :      '++';
DBL_MINUS    :      '--';
RARROW       :      '->';
PHATRARROW   :      '=>';
LSHIFT       :      '<<';
//RSHIFT     :      '>>'; // this is only supported by 2x GT because of: var v<v<int>> a;
//              // which is two templates closing, not a right shift

SCOPE        :      '::';
AT_BANG      :      '@!';
AT_QUEST     :      '@?';
AT_PLUS      :      '@+';
AT_LBRACK    :      '@[';
AUTOINDEX   :      '#' DIGIT;
LBRACK       :      '[';
RBRACK       :      ']';
LCURLY       :      '{';
RCURLY       :      '}';
QM_LPAREN    :      '?(';
LPAREN       :      '(';
RPAREN       :      ')';
AT           :      '@';
AMP          :      '&';
STAR         :      '*';
SLASH        :      '/';
MOD          :      '%';
PLUS         :      '+';
MINUS        :      '-';
SEMI         :      ';';
```

```

COLON      : ':' ;
COMMA      : ',' ;
QM_POINT_STAR : '?.*' ;
QM_POINT   : '?.' ;
QM_LBRACK  : '?[' ;
DOT        : '.' ;
CROSS      : 'x' ;
DIV        : '/' ;
POINT      : '.' ;
EM         : '!' ;
TILDE      : '~' ;
DBL_PIPE   : '||' ;
PIPE       : '|' ;
QM         : '?' ;
HAT        : '^' ;
USCORE     : '_' ;

EQ         : '==' ;
NEQ        : '!=' ;
LTEQ       : '<=' ;
GTEQ       : '>=' ;
LT         : '<' ;
GT         : '>' ;

ASSIGN     : '=' ;
AS_POW     : '**=' ;
AS_MUL     : '*=' ;
AS_SLASH   : '/=' ;
AS_MOD     : '%=' ;
AS_DOT     : '.*=' ;
AS_CROSS   : 'x=' ;
AS_DIV     : '/=' ;
AS_ADD     : '+=' ;
AS_SUB     : '-=' ;
AS_LSH     : '<<=' ;
AS_RSH     : '>>=' ;
AS_AND     : '&=' ;
AS_OR      : '|=' ;
AS_XOR     : '^=' ;

AUTO       : 'auto' ;
VOID       : 'void' ;
BOOL       : 'bool' ;
INT        : 'int' ;
UINT       : 'uint' ;
ISIZE      : 'isize' ;
USIZE      : 'usize' ;
BYTE       : 'byte' ;
CHAR       : 'char' ;
CODEPOINT  : 'codepoint' ;
STRING     : 'string' ;
//HALF     : 'half' ;
FLOAT      : 'float' ;
//DOUBLE    : 'double' ;
//LONGDOUBLE : 'longdouble' ;

I64        : 'i64' ;
I32        : 'i32' ;
I16        : 'i16' ;
I8         : 'i8' ;
U64        : 'u64' ;
U32        : 'u32' ;
U16        : 'u16' ;
U8         : 'u8' ;
B64        : 'b64' ;
B32        : 'b32' ;
B16        : 'b16' ;
B8         : 'b8' ;
F128       : 'f128' ;    // long double prec. float
F64        : 'f64' ;    // double prec. float
F32        : 'f32' ;    // single prec. float
F16        : 'f16' ;    // half prec. float

NS         : 'namespace' ;
MODULE     : 'module' ;
IMPORT     : 'import' ;
VOLATILE   : 'volatile' ;
STABLE     : 'stable' ;
CONST     : 'const' ;
MUTABLE    : 'mutable|mut' ;

```

```

//STATIC      : 'static';
//EXTERN      : 'extern';
PUB           : 'public'|'pub';
PRIV          : 'private'|'priv';
PROT          : 'protected'|'prot';
USING         : 'using';
ALIAS         : 'alias';
UNION         : 'union';
STRUCT        : 'struct';
CLASS         : 'class';
CTOR         : 'ctor';
COPYCTOR      : 'copyctor'|'copy_ctor'|'cctor';
MOVECTOR     : 'movector'|'move_ctor'|'mctor';
DTOR          : 'dtor';
COPYASSIGN    : 'copy=';
MOVEASSIGN    : 'move=';
FUNC          : 'func';
PROC          : 'proc';
METHOD        : 'method';
ENUM          : 'enum';
CONCEPT     : 'concept';
REQUIRES      : 'requires';
PROP          : 'prop';
GET           : 'get';
REFGET        : 'refget';
SET           : 'set';
FIELD         : 'field';
OPERATOR      : 'operator';
VAR           : 'var';
LET           : 'let';
LOOP          : 'loop';
FOR           : 'for';
DO            : 'do';
WHILE         : 'while';
TIMES         : 'times';
IF            : 'if';
ELSE          : 'else';
SWITCH        : 'switch';
DEFAULT       : 'default';
CASE          : 'case';
BREAK         : 'break';
FALL          : 'fall';
RETURN        : 'return';
SIZEOF        : 'sizeof';
NEW           : 'new';
DELETE        : 'delete';
THROW         : 'throw';

ID            : ALPHA_ ALNUM_*;

NUL           : 'null'|'nullptr';
CLASS_LIT     : 'this'|'self'|'base'|'super';
BOOL_LIT      : 'true'|'false';
FLOAT_LIT     : ( DIGIT* '.' DIGIT+ ( [eE] [+-]? DIGIT+ )?
                  | DIGIT+ [eE] [+-]? DIGIT+
                  ) [LfLF]?;
HEX_LIT       : '0x' HEXDIGIT HEXDIGIT_*;
OCT_LIT       : '0o' OCTDIGIT OCTDIGIT_*;
BIN_LIT       : '0b' BINDIGIT BINDIGIT_*;
INTEGER_LIT   : (DIGITNZ DIGIT_*|[0]);

fragment DIGITNZ : [1-9];
fragment DIGIT   : [0-9];
fragment DIGIT_  : [0-9_];
fragment HEXDIGIT : [0-9A-Fa-f];
fragment HEXDIGIT_ : [0-9A-Fa-f_];
fragment OCTDIGIT : [0-7];
fragment OCTDIGIT_ : [0-7_];
fragment BINDIGIT : [0-1];
fragment BINDIGIT_ : [0-1_];
fragment ALPHA_  : [A-Za-z_];
fragment ALNUM_  : [0-9A-Za-z_];

NL           : ( '\\r'| '\\n' )+ -> channel(NEWLINES);
WS           : ( ' '| '\\t' )+ -> skip; // channel(HIDDEN);

```

7.4.2 Grammatik des Parsers

```

parser grammar MyllParser;

options { tokenVocab = MyllLexer; }

comment      :      COMMENT;

// all operators handled by ToOp except mentioned
postOP       :      v=(      '++' | '--' );

// handled by ToPreOp because of collisions
preOP        :      v=(      '++' | '--' | '+' | '-' | '!' | '~' | '*' | '&' );

powOP        :      '*' '*';
multOP       :      v=(      '*' | '/' | '%' | '&' | '.' | 'x' | '÷' );
addOP        :      v=(      '+' | '-' | '^' | '|' ); // split xor and or?
shiftOP      :      '<<' | '>>';

cmpOp        :      '<=>';
orderOP      :      v=(      '<=' | '>=' | '<' | '>' );
equalOP      :      v=(      '==' | '!=' );

andOP        :      '&&';
orOP         :      '||';

nulCoalOP    :      '?:';

memAccOP     :      v=(      '.' | '?.' | '->' );
memAccPtrOP  :      v=(      '.' | '?.' | '->*' );
//memAccOP   :      v=(      '.' | '?.' | '?..' | '->' );
//memAccPtrOP :      v=(      '.' | '?.' | '?..' | '->*' );

// handled by ToAssignOp because of collisions
assignOP     :      '=';
aggrAssignOP :      v=(      '**=' | '*=' | '/=' | '%=' | '&=' | '+==' | 'x=' | '÷=' |
                          '+=' | '-=' | '|=' | '^=' | '<=>' | '>=>' );

lit          :      CLASS_LIT | HEX_LIT | OCT_LIT | BIN_LIT | INTEGER_LIT | FLOAT_LIT | STRING_LIT |
CHAR_LIT | BOOL_LIT | NUL;
wildId       :      AUTOINDEX | USCORE;
id           :      ID;
idOrLit      :      id | lit;

// +++ handled
specialType  :      v=( AUTO | VOID | BOOL );
charType     :      v=( CHAR | CODEPOINT | STRING );
floatingType :      v=( FLOAT | F128 | F64 | F32 | F16 ); // 80 and 96 bit?
binaryType   :      v=( BYTE | B64 | B32 | B16 | B8 );
signedIntType :      v=( INT | ISIZE | I64 | I32 | I16 | I8 );
unsignIntType :      v=( UINT | USIZE | U64 | U32 | U16 | U8 );

qual         :      v=( CONST | MUTABLE | VOLATILE | STABLE );

typePtr      :      qual*
                  ( ptr=( DBL_AMP | AMP | STAR | PTR_TO_ARY )
                  | ary=( AT_LBRACK | LBRACK ) expr? RBRACK )
                  suffix=( EM | PLUS | QM )?;

idTplArgs   :      id tplArgs?;

typespec     :      qual*
                  ( typespecBasic      typePtr*
                  | FUNC typePtr* typespecFunc
                  | typespecNested typePtr* );

typespecBasic :      specialType
                  | charType
                  | floatingType
                  | binaryType
                  | signedIntType
                  | unsignIntType;

typespecFunc  :      funcTypeDef (RARROW typespec)?;

// TODO different order than ScopedExpr
typespecNested :      idTplArgs (SCOPE idTplArgs)*;
typespecsNested :      typespecNested (COMMA typespecNested)* COMMA?; // trailing COMMA here

```

```

really possible?

// --- handled

arg      :      (id COLON)? expr;
args     :      arg (COMMA arg)* COMMA?;
funcCall :      ary=( QM_LPAREN | LPAREN )      args?      RPAREN;
indexCall :      ary=( QM_LBRACK | LBRACK )      args      RBRACK;

param    :      typespec id?;
funcTypeDef :      LPAREN (param (COMMA param)* COMMA?)? RPAREN;

// can't contain expr, will fck up through idTplArgs with multiple templates (e.g. op | from enums)
tplArg    :      lit | typespec;
tplArgs   :      LT tplArg (COMMA tplArg)* COMMA? GT;

tplParams :      LT id (COMMA id)* COMMA? GT;

threeWay :      (orderOP|equalOP)      COLON      expr;

// Tier 3
//cast: nothing = static, ? = dynamic, ! = const & reinterpret
// TODO REMOVE THEM ALL, IN CODE TOO
//
//preOpExpr :      preOP;
//castExpr  :      (MOVE|LPAREN (QM|EM)? typespec RPAREN);
//sizeofExpr :      SIZEOF;
//deleteExpr :      DELETE (ary='['']')?;

// The order here is significant, it determines the operator precedence
expr      :      (idTplArgs      SCOPE)+ idTplArgs      # ScopedExpr
|
|      (
|          postOP
|          funcCall
|          indexCall
|          memAccOP idTplArgs )      # PostExpr
| // can not even be associative without a contained expr
|      NEW      typespec?      funcCall?      # NewExpr
| // inherent <assoc=right>, so no need to tell ANTLR
|      (
|          (      MOVE // parens included
|              LPAREN (QM|EM)? typespec RPAREN)
|          SIZEOF
|          DELETE ( ary='['']' )?
|          preOP
|      )
|      expr      # PreExpr // this visitor

is unused
|      expr      memAccPtrOP      expr      # MemPtrExpr
| <assoc=right>
|      expr      powOP      expr      # PowExpr
|      expr      multOP      expr      # MultExpr
|      expr      addOP      expr      # AddExpr
|      expr      shiftOP      expr      # ShiftExpr
|      expr      cmpOp      expr      # ComparisonExpr
|      expr      orderOP      expr      # RelationExpr
|      expr      equalOP      expr      # EqualityExpr
|      expr      andOP      expr      # AndExpr
|      expr      orOP      expr      # OrExpr
|      expr      nulCoalOP      expr      # NullCoalesceExpr
// TODO: cond and throw were in the same level, test if this still works fine
| <assoc=right>
|      expr      QM expr COLON      expr      # ConditionalExpr
| <assoc=right>
|      expr      DBL_QM threeWay+ (COLON expr)?      # ThreeWayConditionalExpr
| <assoc=right>
|      THROW      expr      # ThrowExpr      // Good or

not?
|      LPAREN expr      RPAREN      # ParenExpr
|      wildId      # WildIdExpr
|      lit      # LiteralExpr
// TODO
|      idTplArgs      # IdTplExpr
;

idAccessor :      id      (LCURLY accessorDef+ RCURLY)?      (ASSIGN expr)?;
idExpr     :      id      (ASSIGN
expr)?;
idAccessors :      idAccessor      (COMMA idAccessor)*      COMMA?;
idExprs     :      idExpr      (COMMA idExpr)*      COMMA?;
typedIdAcors :      typespec      idAccessors      SEMI;

```

```

attribId : id | CONST | FALL | THROW;
attrib   : attribId
        (
            '=' idOrLit
            | '(' idOrLit (COMMA idOrLit)* COMMA? ')'
        )?;
attribBlk : LBRACK attrib (COMMA attrib)* COMMA? RBRACK;

caseStmt : CASE expr (COMMA expr)*
        (
            COLON          levStmt+ (FALL SEMI)?
            | LCURLY       levStmt* (FALL SEMI)? RCURLY
            | PHATRROW     levStmt (FALL SEMI)?);

initList : COLON id funcCall (COMMA id funcCall)* COMMA?;

// is just SEMI as well in levStmt->inStmt
funcBody : PHATRROW expr SEMI
        | levStmt;
accessorDef : attribBlk? a=( PUB | PROT | PRIV )?
        qual* v=( GET | REFGET | SET ) funcBody;
funcDef : id          tplParams?   funcTypeDef (ARROW typespec)?
        (REQUIRES typespecsNested)? // TODO
        funcBody;
opDef : STRING_LIT    tplParams?   funcTypeDef (ARROW typespec)?
        (REQUIRES typespecsNested)? // TODO
        funcBody;
// TODO: can be used in more places
condThen : LPAREN expr RPAREN    levStmt;

// DON'T refer to these in* here, ONLY refer to lev* levels
// ns, class, enum, func, ppp, c/dtor, alias, static
inDecl : NS id (SCOPE id)* SEMI # Namespace // or
better COLON
        | NS id (SCOPE id)* LCURLY levDecl+ RCURLY # Namespace
        | v=( STRUCT | CLASS | UNION ) id tplParams?
          (COLON bases=typespecsNested)?
          (REQUIRES reqs=typespecsNested)? // TODO
          LCURLY levDecl* RCURLY # StructDecl
        | CONCEPT id tplParams? // TODO
          (COLON typespecsNested)?
          LCURLY levDecl* RCURLY # ConceptDecl
        // TODO aspect
        | ENUM id
          (COLON bases=typespecBasic)?
          LCURLY idExprs RCURLY # EnumDecl
        | v=( FUNC | PROC | METHOD )
          ( LCURLY funcDef* RCURLY
            | funcDef ) # FunctionDecl
        | OPERATOR
          ( LCURLY opDef* RCURLY
            | opDef ) # OpDecl
// class only:
        | v=( PUB | PROT | PRIV ) COLON # AccessMod
        | v=( CTOR | COPYCTOR | MOVECTOR )
          funcTypeDef initList? (SEMI | levStmt) # CtorDecl
        | DTOR LPAREN RPAREN (SEMI | levStmt) # DtorDecl
        ;

// using, var, const: these are both Stmt and Decl
inAnyStmt : USING          typespecsNested SEMI # Using
        | ALIAS id ASSIGN typespec SEMI # AliasDecl
        | v=( VAR | FIELD | CONST | LET )
          ( LCURLY typedIdAcors* RCURLY
            | typedIdAcors ) # VariableDecl
        ;

inStmt : SEMI # EmptyStmt
        | LCURLY levStmt* RCURLY # BlockStmt
        | RETURN expr? SEMI # ReturnStmt
        | THROW expr SEMI # ThrowStmt
        | BREAK INTEGER_LIT? SEMI # BreakStmt
        | IF
          (ELSE IF condThen)* // helps with formatting properly and de-nesting the AST
          (ELSE levStmt)? # IfStmt
        | SWITCH LPAREN expr RPAREN LCURLY
          caseStmt+ (DEFAULT levStmt+)? RCURLY # SwitchStmt
        | LOOP levStmt # LoopStmt
        | FOR LPAREN levStmt
          expr SEMI // TODO: add the syntax: for( a : b )

```



```

        expr RPAREN      levStmt
        (ELSE             levStmt)?      # ForStmt
    |   WHILE             condThen
        (ELSE             levStmt)?      # WhileStmt
    |   DO                levStmt
        WHILE LPAREN expr RPAREN          # DowhileStmt
    |   DO expr TIMES      id?      levStmt # TimesStmt
    |   expr (assignOP      expr)+ SEMI   # MultiAssignStmt
    |   expr aggrAssignOP  expr SEMI      # AggrAssignStmt
    |   expr SEMI          # ExpressionStmt
    ;

// ONLY refer to these lev* levels, NOT the in*
levDecl : attribBlk LCURLY levDecl+ RCURLY # AttribDeclBlock // must be in here, since
it MUST have an attrib block
// | attribBlk COLON // everything needs an antonym to make this work
levStmt : attribBlk? ( inAnyStmt | inDecl ) # AttribDecl;
        attribBlk? ( inAnyStmt | inStmt ) # AttribStmt;

module : MODULE id SEMI;
imports : IMPORT id (COMMA id)* COMMA? SEMI;

prog : module? imports* levDecl+;

```