Unicode & Emoji

Mark Davis

President & Co-founder, Unicode Consortium
@mark_e_davis

https://goo.gl/Lff11S

What is the Unicode Consortium?

Enable everybody,

speaking every language on the Earth,

to be able to use their language

on computers and smartphones

Levels of Language Support

Core	Characters, Fonts, Keyboards
Content	Sorting, Numbers, Dates, Units, Currency, …
UI	Translated UI, Help text, …
NLP	Predictive typing, Voice input / output, OCR, Entity recognition, Handwriting input, …

Full Members (Voting)




Institutional Members (Voting)

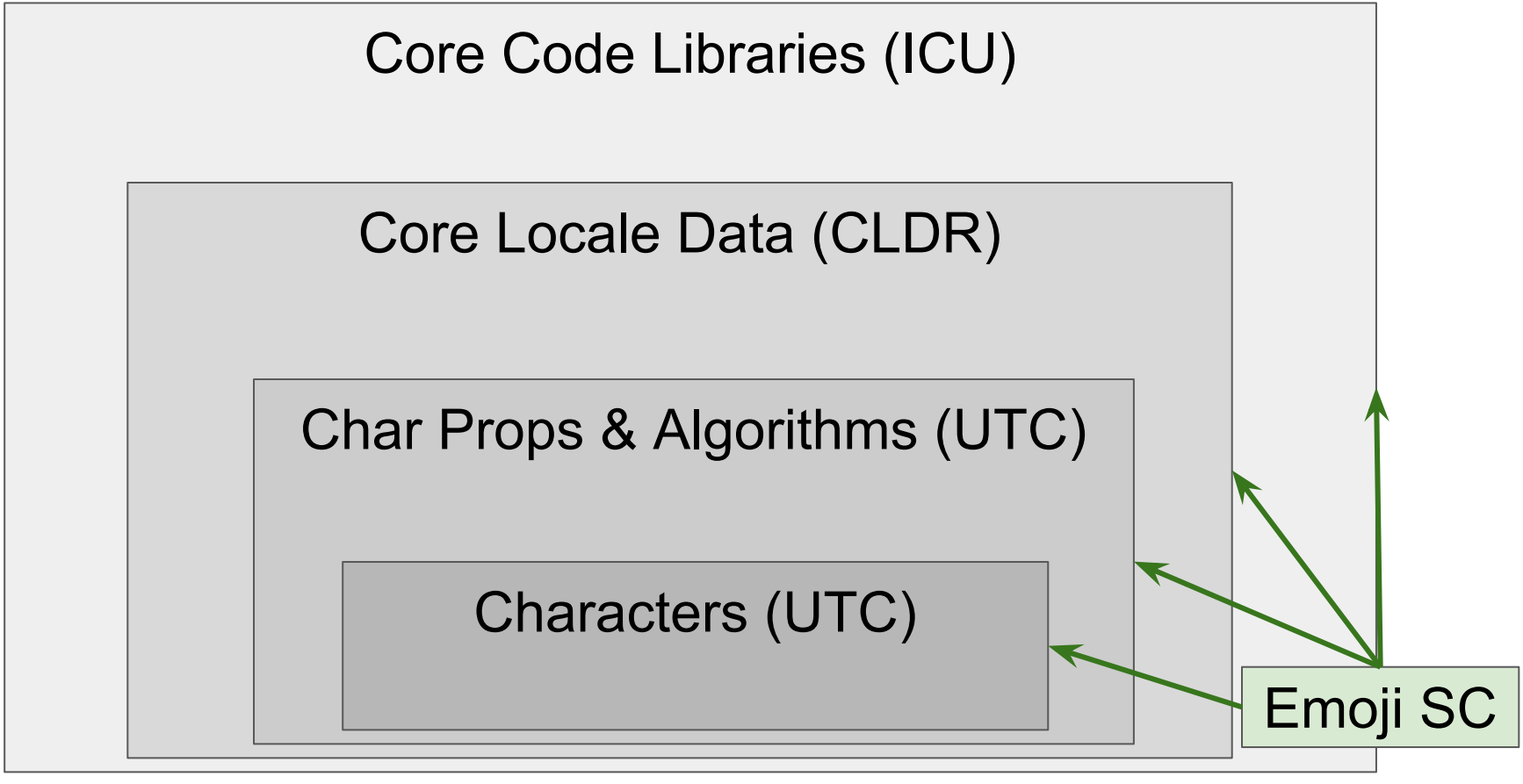


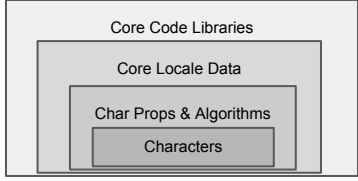
Supporting Members (Voting)

Associate Members


member-supported, non-profit

Unicode Consortium





What is a Unicode Character?

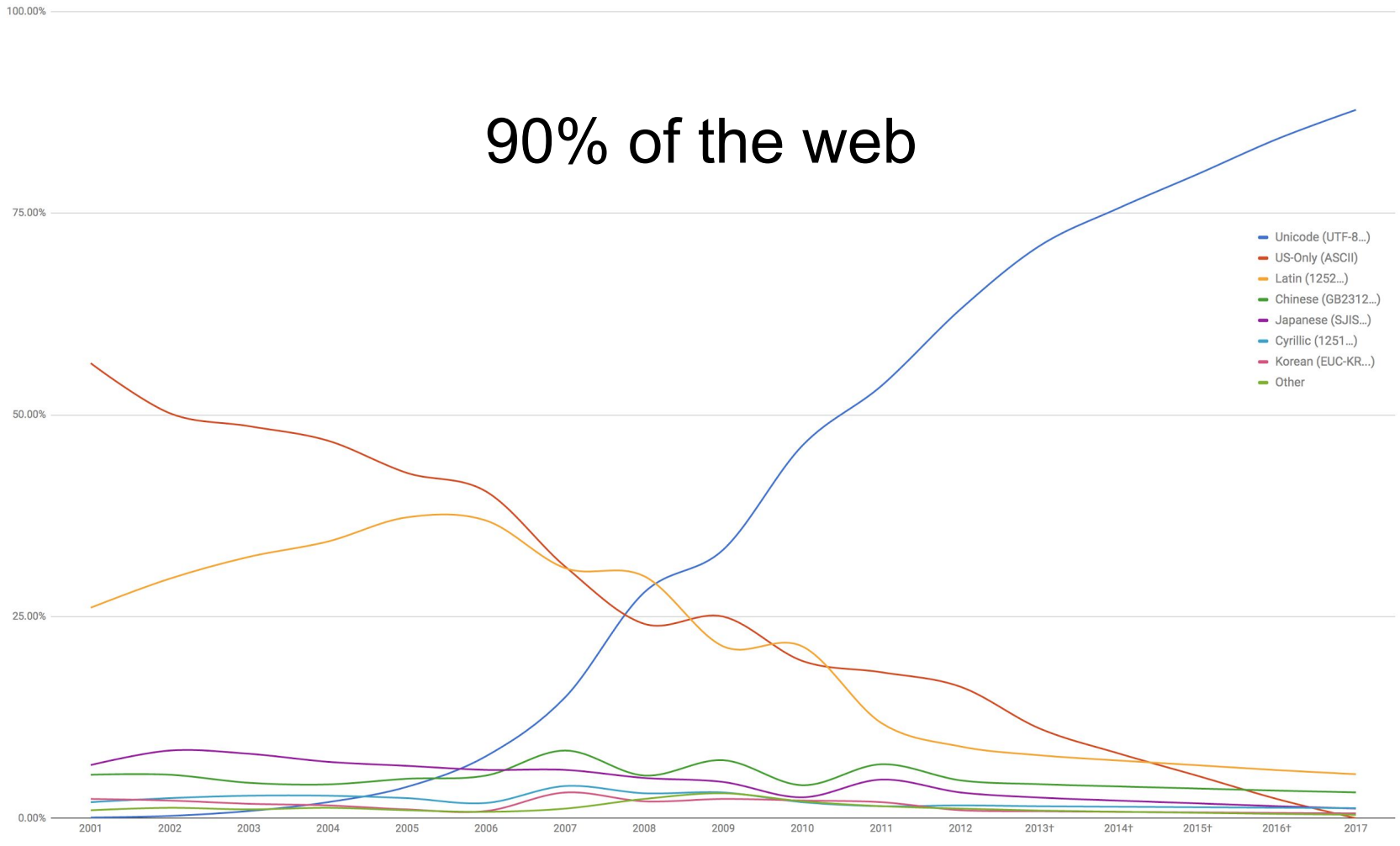
On a laptop, server, or mobile phone

every character you type,

every character you see

is Unicode

90% of the web



googleblog.blogspot.ch/2012/02/unicode-over-60-percent-of-web.html

— graph updated to 2017

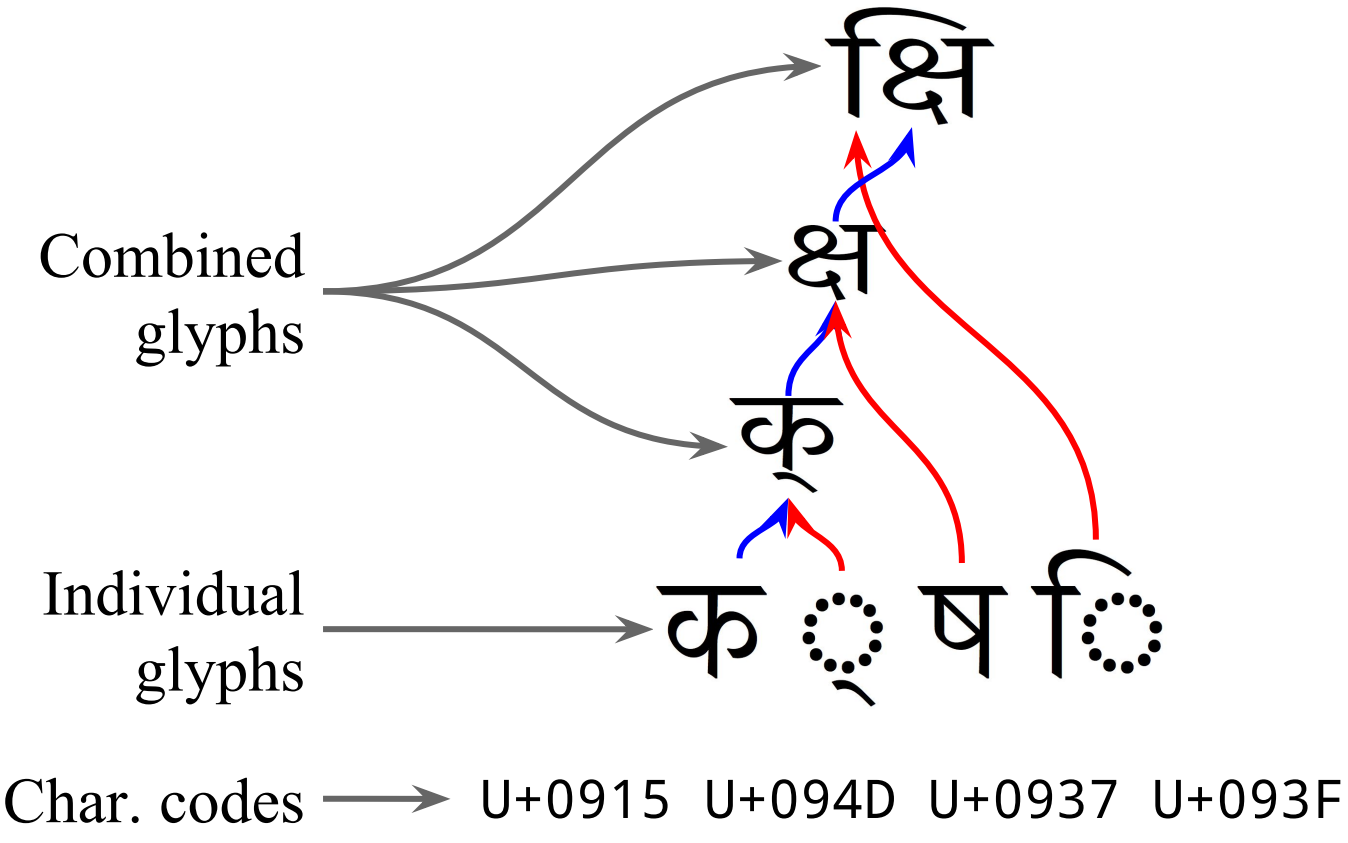
Not as simple as you think

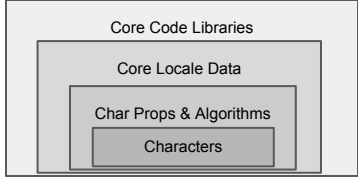
Combined glyphs → fish

Individual glyphs → f i s h

Char. codes → U+0066 U+0069 U+0073 U+0068

Not as simple as you think





Why Properties?

Full name

Zoë Straub

Illegal Name

Before

```
if
   ‘a’ ≤ X AND X ≤ ‘z’

then handle(X)
```

After

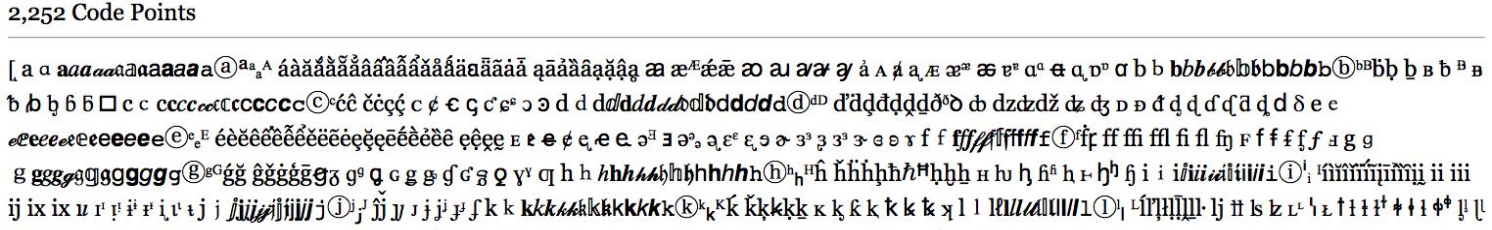
```
If
 isLowercase(X)

then handle(X)
```

isLowerCase(…)

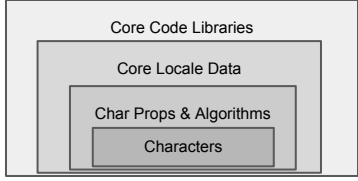
Data in updated in each version of Unicode:

2,252 Code Points



Unicode Properties

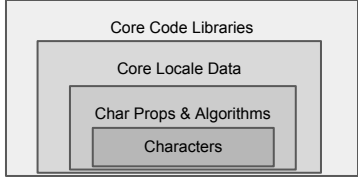
General	Case	Normalization	Shaping and Rendering
Name	Uppercase	Canonical_Combining_Class	Join_Control
Name_Alias	Lowercase	Decomposition_Mapping	Joining_Group
Block	Lowercase_Mapping	Composition_Exclusion	Joining_Type
Age	Titlecase_Mapping	Full_Composition_Exclusion	Line_Break
General_Category	Uppercase_Mapping	Decomposition_Type	Grapheme_Cluster_Break
Script	Case_Folding	NFC_Quick_Check	Sentence_Break
Script_Extensions	Simple_Lowercase_Mapping	NFKC_Quick_Check	Word_Break
White_Space	Simple_Titlecase_Mapping	NFD_Quick_Check	East_Asian_Width
Alphabetic	Simple_Uppercase_Mapping	NFKD_Quick_Check	Prepended_Concatenation_Mark
Hangul_Syllable_Type	Simple_Case_Folding	NFKC_Casefold	**Bidirectional**
Noncharacter_Code_Point	Soft_Dotted	Changes_When_NFKC_Casefolded	Bidi_Class
Default_Ignorable_Code_Point	Cased	**Miscellaneous**	Bidi_Control
Deprecated	Case_Ignorable	Math	Bidi_Mirrored
Logical_Order_Exception	Changes_When_Lowercased	Quotation_Mark	Bidi_Mirroring_Glyph
Variation_Selector	Changes_When_Uppercased	Dash	Bidi_Paired_Bracket
Identifiers	Changes_When_Titlecased	Sentence_Terminal	Bidi_Paired_Bracket_Type
ID_Continue	Changes_When_Casefolded	Terminal_Punctuation	**CJK**
ID_Start	Changes_When_Casemapped	Diacritic	Ideographic
XID_Continue	**Numeric**	Extender	Unified_Ideograph
XID_Start	Numeric_Value	Grapheme_Base	Radical
Pattern_Syntax	Numeric_Type	Grapheme_Extend	IDS_Binary_Operator
Pattern_White_Space	Hex_Digit	Indic_Positional_Category	IDS_Trinary_Operator
	ASCII_Hex_Digit	Indic_Syllabic_Category	Unicode_Radical_Stroke



Unicode Algorithms

Unicode detection, conversion	Script detection
Normalization, equivalence	Case detection, conversion
Collation (sorting)	Segmentation
Regular expressions	Unicode domain names
Bidirectional text	Security mechanisms
Identifier parsing	Emoji validity
Shaping (Arabic,…)	Vertical orientation (CJK)
…	

Language-Specific Data (CLDR)

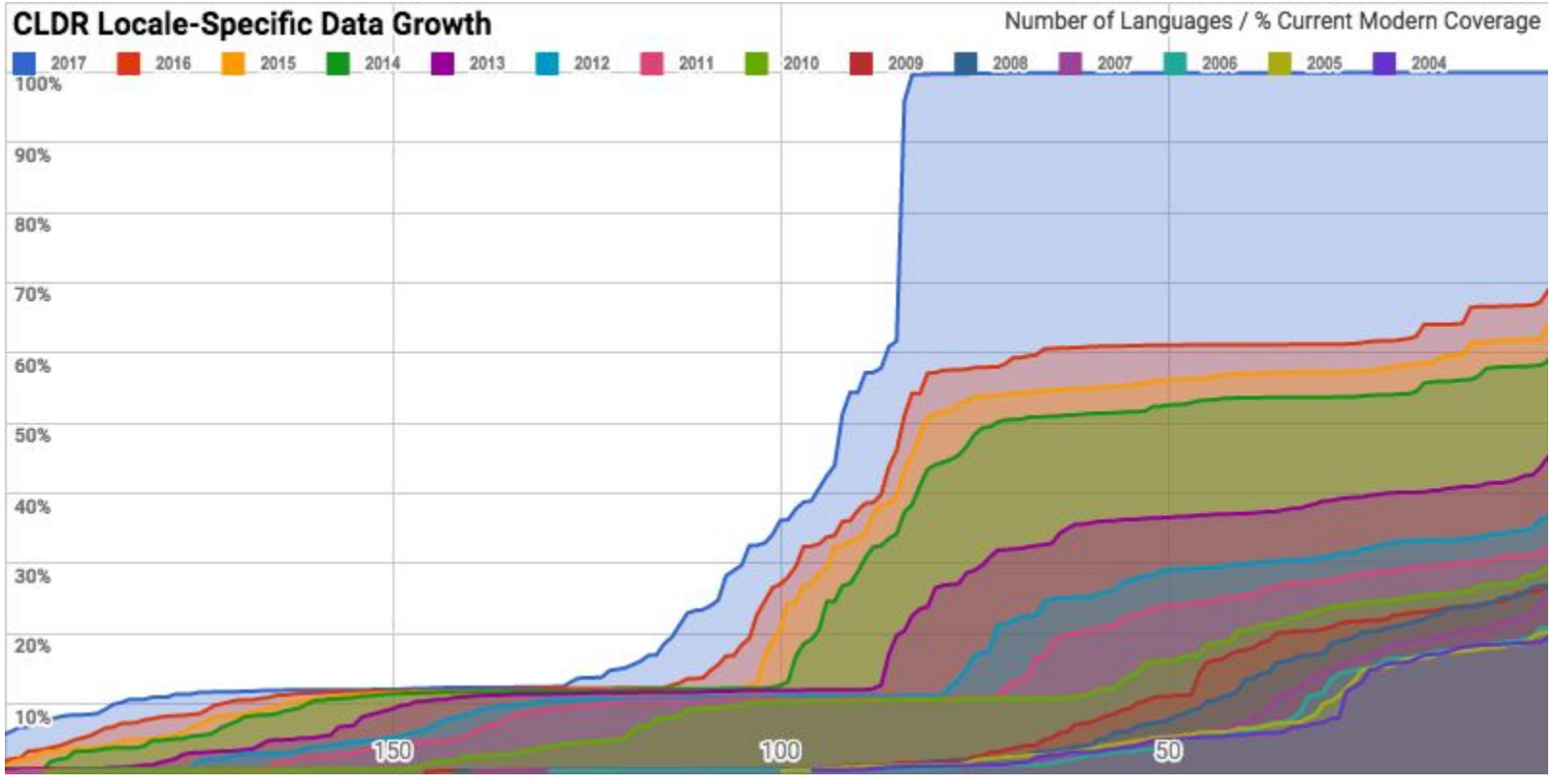


Characters:	ä, ø,…
Sorting	ä < b vs ä > z
Dates & times	Dec 3–5; last Tuesday
Day periods, timezones	8:00 → morning1; 13:52 Italy Time
Numbers, units, currencies	1.234; $3.5M; 123 Kanadische Dollar
Lists, ranges, compact forms	Dec 3–5; 3.5–4.2kg; 3.5B
Names of languages, scripts,…	FR → 法语; Cyrl → Kyrillisch; 419 → Südamerika
Character labels, names,…	□ → Arrows; ♉ → Taurus; bull | ox | zodiac
Transforms	Путин → Putin; Трамп → Trump
Spell out numbers	25 → twenty-five
Plurals (cardinals, ordinals)	1 book, 2 books; 1st book; 1–2 books
Keyboards	C03 → д; C03+*caps* → Д
Locale matching/normalization, Region info (currencies, containment, …) …	

CLDR Contributors

A S Alam (অমলপরীত সিঁথি আলম), Abdirahman Ali Mohamed, Abdullah Hassaan Al-Anssary (عبدالله حسان الأنصاري), Abduqadir Abliz(نابلىز قادىر), Adam Peller, Adam Richeimer, Addisu Alemneh Tesfaye, Dr. Adebisi Aromolaran, Adélaïde Bodson, Adham Kurbanov, Admir Qose, Adriana Adarve, Agustín Da Fieno Delucchi, Ahmad Farrag (أحمد فرّاج), Ahmed Essam El Din (أحمد عصام الدين), Ahmed Gad (أحمد جاد), Ahmed Najib Biabani (احمد نجیب بیابانی), Åke Persson, Akio Kido, Alan Liu, Alberto Barbati, Alberto Escudero-Pascual, Alberto Picon, Alejandra Canoura Pérez, Aleksandr Andreev (Александр Андреев), Alexander Buloichik (Аляксандр Булойчык), Alexander Kachur (Александр Качур), Alexandre Moréteau, Alexey Proskuryakov (Алексей Проскуряков), Alexis Cheng (鄭宗堯), Amandeep Singh Saini (অমনদীপ সিঁঘ সৈনী), Amir E. Aharoni (אמיר א' אהרוני), Amir El-Sharqawy (أمير الشرقاوي), Amir Sarabadani (امیر سرابادانی), Amr Emad El Din (عمرو عماد الدين), Amy Kilby, Ana Kušen, Anastasia Golikova (Анастасия Голикова), André Malafaya Baptista, André Stryger, Andrea Decorte, Andreas Nieckele, Andy Heninger, Aneta Risteska, Angelo Dalli, Anil Goyal (अनिल गोयल), Aniza Borhan, Anna Sixkiller, Anna Taranova (Анна Таранова), Anna Zaytseva (Анна Зайцева), Annelize Swart, Annette Wiibroe Lorentzen, Anoop.M.R (അനൂപ്.എം.ആർ), Anthony Deign (அன்தோனி டெய்ன்), Antoine Leca, Aram Palyan (Արամ Պալյան), Artur Klauser, Aruna Kumar KR (ಅರುಣ ಕುಮಾರ್ ಕೆ ಆರ್), Arvinder Singh Kang (ਅਰਵਿੰਦਰ ਸਿੰਘ ਕੰਗ), Ashok Reddy (అశోక్ రెడ్డి), Ashraf-ul-Kabir (اشرف الكبير), Atakan Özdemir, Atso Puronen, Audun Lona, Aungthu Schlenker, Aurelian George Dumanovschi, Avery Chan, Aviel Boag (אביאל בואג), Ayushya Kishore Devmurari (આયુષ્ય કિશોરભાઇ દેવમુરારી), Bakytgul Salykhova (Бакытгул Салыхова), Bal Krishna Bal (बालकृष्ण बल), Balaji Natarajan (பாலாஜி நடராஜன்), Baldev Soor, Barka Elijah Gituwa, Bayu Saputra, Behdad Esfahbod (بهداد اصفهبد), Ben Black Bear, Berend Ytsma, Bert Jickty (Jacob E. Alexandrov), Bhavna Mishra (भावना मिश्रा), Bhupal Sapkota (भूपाल सापकोटा), Biligsaikhan Batjargal (Батжаргалын Билигсайхан), Bjørn Hope, Blanca Amoroso, Bojan Jankuloski (Бојан Јанкулоски), Boris Matveev (Борис Матвеев), Britta Sinks, Calvin Lee (李灝), Carine Awada (كارين عواضة), Carlos Felipe Anonorsi, Carlos Koch, Caroline Jacob, Cătălin Boaru, Catarina Simoes, Catherine Christaki (Κατερίνα Χριστάκη), Celalettin Penbe, Charitha Madusanka (චරිත මධුසංඛ), Charles Pau (鮑景超), Charles Riley, Charlotta af Hällström-Reijonen, Chiemerie Ogenyi-Benjamin, Chloe Chan (陳浩敏), Cho Akum Ndi, Chookij Vanatham (ชูกิจ วนาธรรม), Chorobek Saadanbekov, Chris Fleizach, Chris Hansten, Chris Leonard, Chris McDowell, Christina El hayek (كريستينا الحايك), Christine Guthry, Christopher Fynn (ཀྲིས་ཊོཕར་ཧྥིན་ནི།), Cibu C. J. (സിബു സി.ജെ.), Claire Ho (賀靜蘭), Clara Bozzo Closas, Constantina Yiokari, Craig Cornelius (ᏣᎳ), Cristiana Coblis, Cristina Gonçalves, Dacijus Turauskas, Dafydd Tomos, Dagmar Merlino, Dale Schultz, Damien Alexandre, Damjan Zorc, Dan Dascalescu, Danh Hong (ฮ้าญ์ ฮ้าฝ), Danica Milošević (Даница Милошевић), Daniel Bruhn, Daniel Polimac, Daniel U. Thibault, Daniel Wojciechowski, Daniel Yacob, Daniela Slankamenac, Dary Mihova (Дари Михова), David Luo (罗凌), David Petit, Deborah Goldsmith, Delyth Prys, Demi Rumble (Дэлгэрмаа Рамбл), Denis Jacquerye, Dennis Sixkiller, Dewi Bryn Jones, Dharma Adhivijaya Gapin, Diana Abdrakhmanova, Dick Veerhar, Dilip Vegad (દિલીપ વેગડ), Dilmurat Abduqayyum (نابۇقەييۇم), Divon Lan (לאן, דיבון), Djordje Vasiljevic (Ђорђе Васиљевић), Dmitry Molodyk, Đoàn Anh Đức, Dominic Hung (洪淦輝), Dong-hyun Kim (김동현), Doug Ewell, Doug Felt, Dov Pearl, Dwayne Bailey, Ed Jumper, Eddy Petrişor, Edwin Otieno Makori, Efthymia Lixourgioti (Ευθυμία Λιξούργιώτη), Eike Rathke, Eirikur Heðinsson Mortensen, Ekkapol Meksuwan (เอกพล เมฆสุวรรณ), Elena Ivanova (Елена Иванова), Eleri James, Eli Anne Budal, Elias E. Haddad (الياس عيد حداد), Elias Mossholm, Elisabete Ferreira, Elisabeth Alster, Elisardo Juncal Muñiz, Elliott Hughes, Elsebeth Flarup, Elva Bian (边文静), Elvana Tufa, Enol Puente Suárez, Erdal Ronahi, Eric Mader, Eric Muller, Erkki Kolehmainen, Ernest v.d. Boogaard, Ernesto Castellanos Lozano, Eva Albertsdotter Enblom, Evdoxia Renta, Fagelot Solofonirina, Fayeq Oweis (فائق عويس), Federico Leva, Firitia Velt, Firoj Alam (फिरोज आलम), Flora Gu, Focko Dijksterhuis, Francisco Javier Rodríguez Arias, Dr. Frans Wijma, Fredrik Roubert, Fredrik Stenshamn, Friedel Wolff, Fulya Becer, Gabrielle Blanc, Gao Hailin(高海林), Gediminas Vaitkevičius, George Rhoten, Georgiana Raluca Lazăr, Gerson Keiti Motoyama, Ghasem Kiani (قاسم کیانی مقدم), Gheyret Toxti Kenji (غەيرەت توختى كەنجى), Gilbert Adjoyi, Gion-Andri Cantieni, Girish G (ಗಿರೀಶ್ ಜಿ), Godwin, Gregory Huczynski, Grishma Shah (ગ્રીષ્મા શાહ), Gruff Prys, Gunnar Hjalmarsson, Hamza Foziljonov, Han Keuper, Hank Wang, Hanna Liljeblad, Hardeep Kaur (ਹਰਦੀਪ ਕੌਰ), Hari Prasad Kafley, Harry Gao, Håvard Hjulstad, Helena Shih Chapman (施尚玲), Hélène Guillerme, Heli Hägg, Helle Hartmann Jensen, Helle Madsen, Hendrayatna Tafianoto, Hideki Hiura, Ian Parsons, Idesh Batkhuu, Igor Viarheichyk (Ігар Вяргейчык), Ihar Mahaniok (Ігар Маханёк), Ilkin Huseynov, Illimar Tambek, Ilyas Bakirov (Ильяс Бакиров), Iñigo Varela, Irina Chausova, Irina Parshina, Iroro Orife, Isabel Saval, Ivailo Djilianov (Ивайло Джилянов), Ivana Jelisavcic (Ивана Јелисавчић), J Meenakshi (மீனாக்ஷி), Jaap Roos, Jacobo Tarrío Barreiro, Jae-in Shim (심재인), Jaganadh.G (ജഗന്നാഥ്.ജി), Jakub Truschka, Jama Musse Jama, James Ballentine, Jan Lána, Jan Ullrich, Jana Bedaňová, Jana Vořechovská, Janaki Christina (ஜானகி கிறிஸ்டினா), Janny Chen (陈敏敏), Jarkko Hietaniemi, Jarosław Włodarczyk, Jason Currie, Javier Sola, Jean-Jacques Enser, Jeff Edwards, Jefferson Bien-Aime, Jehan, Jelena Oertel, Jelena Pjesivac-Grbovic (Јелена Пјешивац-Грбовић), Jennifer Chye, Jeno Demeczky, Jeow Li Huan, Jeroen Ruigrok van der Werven, Jinia Dutta (জিনিয়া দত্ত), Joan Montané, John Emmons, Jolanta Szczerba, Jon Harald Søby, Jonathan Lai, Jonathan Pool, Jonathan Rosenne (יונתן רוזן), Jordi Mallach, Jørgen Christian Wind Nielsen, Jose Bernardo Soto Insua, Josep M. Closa, Joseph Erb, Judit Subirachs Guix, Jung Hyoun Yi (이정현), Jungshik Shin (신정식, 申政浣), Junko Sakakibara (榊原 淳子), Kalamalini Sundararajan, Kamaljit Singh(ਕਮਲਜੀਤ ਸਿੰਘ), Karan Misra (करन मिश्र), Karin Colsman, Katja Hoisko, kazède king, Keghani Kouzoujian (Գեղանի Գուզուճեան), Keith Stribley, Kent Karlsson, Kentaroh Noji (野地 健太郎), Keshan Sodimana (කේෂාන් සෝදිමාන), Kevin Scannell, Khaled Hosny (خالد حسني), Khamphay Inthara (ຄຳພາຍ ອິນທະຣາ), Kiala Ntona, Dr Kipkoeech araap Sambu, Kirill Buryak, Klaas Ruppel, Klean Xhelilaj, Kresna Luginawati, Kudzanai Chitsungo, Kyoung-Jun Min (민경준), Lara Vázquez Cao, Laurențiu Iancu, Lauris Bukšis-Haberkorns, Lazar Milicevic Lee Collins, Leila Nayernia (لیلا نیرنیا), Leo Hanuš, Leonardo Brondani Schenkel, Li Ming Yeung (李名揚), Li Xianxiang(黎先翔), Liath Noy (ליאת נוי), Liesbeth Tjemmes, Lilavati Vaidya (लीलावती वैद्य), Lilia Borissova (Лилия Борисова), Lipika Yadav (लिपिका यादव), Livia Soardi, Luke Sandberg, Lynn Lin (林蓓蓓), Maheswari Raju (மகேஸ்வரி ராஜூ), Maiko Nagae Tasaki (田崎 麻衣子), Maksim Rudolf, Malte Wedel, Manfred Kiefer, Mangat veer Sagar (Mangat), Manjistha Bhattacharya (মঞ্জিষ্ঠা ভট্টাচার্য), Manoj MK (മനോജ് എം.കെ), Mansur Kabirov (Мансур Кабиров), Marc Liyanage, Marc-André Séguin, Marcellus van Lohuizen, Márcio Ferreira, Maria Luna, Maris Vilks, Marius Swart, Marjan Shirazi (مرجان شیرازی), Mark Davis, Mark Garrett, Markus Scherer, Martin Brümmer, Martin Jansche, Martin Raymond, Martti Roslakka, Masanori Goto (後藤 正徳), Mati Allouche (מתתיהו אלוש), Matt Hemmings, Matthias Dieter Wallnöfer, Max Heller, Michael Bauer, Michael Everson, Mícheál John Ó Meachair, Michael Rohan, Michael Twomey, Michaela Boteva (Михаела Ботева), Michal Dadan, Michele Locati, Miguel A. Jimenez, Miguel da Fonseca, Miguel Sousa, Miha Rus, Mihai Niță, Mike Tardif, Miklós Surján, Mikołaj Zalewski, Milind Wanjari, Mine Halıcı, Minh Phireak (មុឺញ ភីរក្ស), Mirela Škarica, Mitra Fathollahpour (میترا فتح اله پور), Mohd Mansoor Alam (محمد منصور عالم), Morteza Yazdani (مرتضی یزدانی), Mthokozisi Moses Dlalisa, Muthu Nedumaran (முத்து நெடுமாறன்), Myriam Bedmar Martín, Myron Nechypor, Nadee Fernando, Naga HV Bhuthi (నాగ హరీష్ వరధ్), Naila Noah (نائلة نوح), Najaf Raza (نجف رضا), Namgay Thinley (ནམ་རྒྱལ་གཙུག་ཕྲེལ།), Nanette Stensholt, Natalia Conovca, Natalie Davitashvili (ნატალია დავითაშვილი), Natia Gabrichidze (ნათია გაბრიჩიძე), Nattapong Sirilappanich (ณัฐพงษ์ ศิริลาภพานิช), Naushina Qamar, Neda Armanfar (ندا آرمانفر), Neel Kamal Thakur (नील कमल ठाकुर), Neelima Talwar (నీలిమా తల్వార్), Dr. Neil Morton, Nemo Semret (ነሞ ሥምረት), Ngwe TUN (ငွေထွန်း), Nicha Kumchokpattaraporn (ณิชชา กำโชคภัทรพร), Nicholas Shanks, Nicole Oberholzer, Nigel Rodgers, Nijolė Miknevičienė, Nik Kalach, Niklas Laxström, Nikola Bogdanović, Nina Åberg, Nkanyiso Mazibuko, Nontokozo Mbuyisa, Noshrevan Chkhaidze (ნოშრე ჩხაიძე), Nova Patch, Nuria Calafell Cardenal, Núria Juhera Bou, Oktay Coskun, Ole Kristian Julseth, Ole Laursen, Omi Azad (অমি আজাদ), Orly González Kahn, Oscar D.P. Triscon, Oscar Hernandez Quiles, P Mett (פרץ מטעט), Pablo Muñoz Sánchez, Pablo R. Estrada, Paloma Kreischer, Pamela Ávalos Parker, Pantelis Assimakopoulos, Parth Gaglani, Pasindu De Silva (පසිඳු දේ සිල්වා), Patrick Andries, Patroba Anyona, Paul G. Hackett, Paulo José, Pavol Haviernik, Pawel Dyda, Peter Edberg, Peter Linsley, Peter Nugent, Petr Kadlec, Petru Haivei, Philippe Verdy, Philips Benjamin, Phyllis Edwards, Pia Nanny, Pravin Jain (પ્રવીણ જૈન), Priya Bhawalkar (प्रिया भवाळकर), Puri Rodríguez Díaz, Purodha Blissenbach, Rachael Ndichu, Rachel Kremer (רחלי קרמר), Radek Podolski, Radi Pipev (Ради Пипев), Radosław Moszczyński, Radosław Niewiarowicz, Rafael Soares Mazzeu da Silva, Raghuram Viswanadha (రఘురామ్ విశ్వనాధ), Raj Neettiyath (രാജ് നീട്ടിയത്ത്), Dr. Rajesh Kumar (डॉ. राजेश कुमार), Rajesh Ranjan (राजेश रंजन), Ramon Casha, Raymond Wainman, Reiko Saito (斎藤 玲子), Riad Khaled Hallab (رياض خالد حلاب), Rici Lake, Rick McGowan, Rob Kiely, Robert Barrasso, Robert Muir, Robert Ribnitz, Roberto Bahamonde A., Robin van der Vliet, Rocío Román López, Roey Saban (רועי סבן), Romaine, Roozbeh Pournader (روزبه پوردانر), Rosidah Mohamad, Rossi Ignatova (Роси Игнатова), Roumen Botev (Румен Ботев), Roy Boney, Ruben Zakharov (Рубен Захаров), Ruby Chiu (邱莉婷), Rustem Arzymbetov, Ruth Layús Quiroga, S. M. Murtoza Habib (এস. এম. মুর্তজা হাবিব), Sajith VK, Saku Airila, Salome Nduku, Samir Kumar Mondal (সমীর কুমার মন্ডল), Samphan Raruenrom (สัมพันธ์ ระรื่นรมย์), Samuel Murray, Sandra O'Donnell, Saniye Boran, Santhosh Thottingal (സന്തോഷ് തോട്ടിങ്ങൽ), Sara Migliori, Sara Santa Clara, Saran Mahasupap (ศรัณย์ มหาสุภาพ), Saravanan (சரவணன்), Sarika Nene (सारिका नेने), Dr. Sarmad Hussain [CRULP (مرکز تحقیقات اردو)], Saroj Kumar Dhakal (सरोज कुमार ढकाल), Sasan Banava (ساسان بانوا), Sascha Brawer, Sasha Maric (Саша Марић), Saurabh Kudesia (सौरभ कुदेशिया), Séamus Ó Ciardhuáin, Sebastian Hellmann, Sedric Motha, Seeun Lee (이세은), Semir Mehadzic, Serena Xiao (肖顺敏), Sergey Karavay (Сергей Каравай), Serpil Zengin Akkaya, Shankar Prasad M V, Shaopeng Jia (贾少鹏), Shawn Steele, Shervin Afshar (شروین افشار), Shi Zhao (时昭), Shihabudheen. P.M (ഷിഹാബുദ്ദീന്.പി.എം), Shinichiro Ishii (石井 信一郎), Shu Bei (舒蓓), Siddaarth Shanmugam (சித்தார்த் சண்முகம்), Silvia Rădulescu, Simon Dean, Simone Lwin, Simone Nicolo, Simos Xenitellis (Σίμος Ξενιτέλλης), Sivanesan Subramaniam, Sivanesh Rajendran (சிவனேஷ் இராஜேந்திரன்), Sivaraj Doddannan, Soe Min Mark, Solomon Gizaw (ሰለሞን ግዛው), Somil Mishra (सोमिल मिश्रा), Somnath Chandra, Sonia Mokrani, Søren Thing Pedersen, Sorin Ionuț Sbârnea, Soslan Khubulov, Sreyleap Lay, Stanislav Pecha, Stanislaw Kulikowski, Stefano Riccardo Rizzardi, Steingrímur Árnason, Sten Thaning, Stepan Herman (Степан Герман), Steven R. Loomis, Su Liu (劉速), Sudhakar Chandrasekharan, Sumesh Chandran V.R (സുമേഷ് ചന്ദ്രന്.വി.ആർ), Sumesh Menon K (സുമേഷ് മേനോൻ കെ), Suresh Chandrasekharan (സുരേഷ് ചന്ദ്രശേഖരൻ), Susan Maria Howard, Sushil Kahat (सुशील कहात), Sushrut Jalukar (सुश्रुत जळूकर), Sussu Laaksonen, Suwit Srivilairith (สุวิทย์ ศรีวิไลฤทธิ์), Suzy Wang, Sverrir Á. Berg, Swaran Lata, Tarek Abou Aly (طارق أبو علي), Tatsuo Sekine (関根 達夫), Teppo Turtiainen, Tero Aalto, Terry Chrysopoulos (Λευτέρης Χρυσόπουλος), Tesfayohannes Tareke (ተስፋዮሃንስ ታረቀ), Thammika Namnakprasert (ธรรมิกา นามนาค ประเสริฐ), Thanavorakit Kounthawatphinyo (ทะนะวอ็ะกิต กุนทะขะวัดพินโย), That Naing Htoo (သက်နိုင်ထူး), Theresa Hieatt-Smith, Therese Svärdsten Nymans, Thomas Dujardin, Thomas Pace, Thomas Ryan, Thrine Bjerknes, Tilek Mamutov (Тилек Мамутов), Tim Brandall, Timothy Stranex, Toliño, Tom Boerger, Tom Garland, Tomas Molnar, Tomasz W. Kozłowski, Travis Keep, Troy Korjuslommi, Truc Tran (Trần Văn Trúc), Umesh Nair (ഉമേഷ് നായർ), Vachagan Gratian, Vadim Sozykin (Вадим Созыкин), Valeriy Perunov (Валерий Перунов), Vardan Grigoryan, Vasyl Shumskyi, Veeven (වීවන්), Vena Zhu (朱为花), Vera Poon, Vicenç Riullop, Vipin Pandit (विपिन पंडित), Virginia Barranco González, Viswa Prabha (വിശ്വപ്രഭ), Vitaliy Shipitsyn (Виталий Шипицын), Vithaya Vonglat, Vladimir Weinstein (Владимир Вайнштейн), Vladislav Zavyalov (Владислав Завьялов), Vytautas Rekus, Waldir Pimenta, Walter Keutgen, Waris Abdukerim Janbaz (وارس نابۇزكېرىم جانباز), Weilin Zhang (张卫琳), William Gan (甘玮亚), William J. Sullivan, Witold Jarnicki,

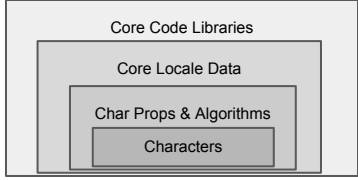
CLDR Growth


CLDR Locale-Specific Data Growth
Number of Languages / % Current Modern Coverage
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
150
100
50

Survey Tool

🐼 `-name`	panda face	✔	**Pandagesicht**
🐾 `-name`	paw prints	✔	**Tatzenabdrücke**
Animal-Bird			
🦃 `-name`	turkey	✔	**Truthahn**
🐔 `-name`	chicken	✔	**Huhn**
🐓 `-name`	rooster	✔	**Hahn**
🐣 `-name`	hatching chick	✔	**schlüpfendes Küken**
🐤 `-name`	baby chick	✔	**Küken**
🐥 `-name`	front-facing baby chick	✔	**Küken in Frontansicht**
🐦 `-name`	bird	✔	**Vogel**
🐧 `-name`	penguin	✔	**Pinguin**

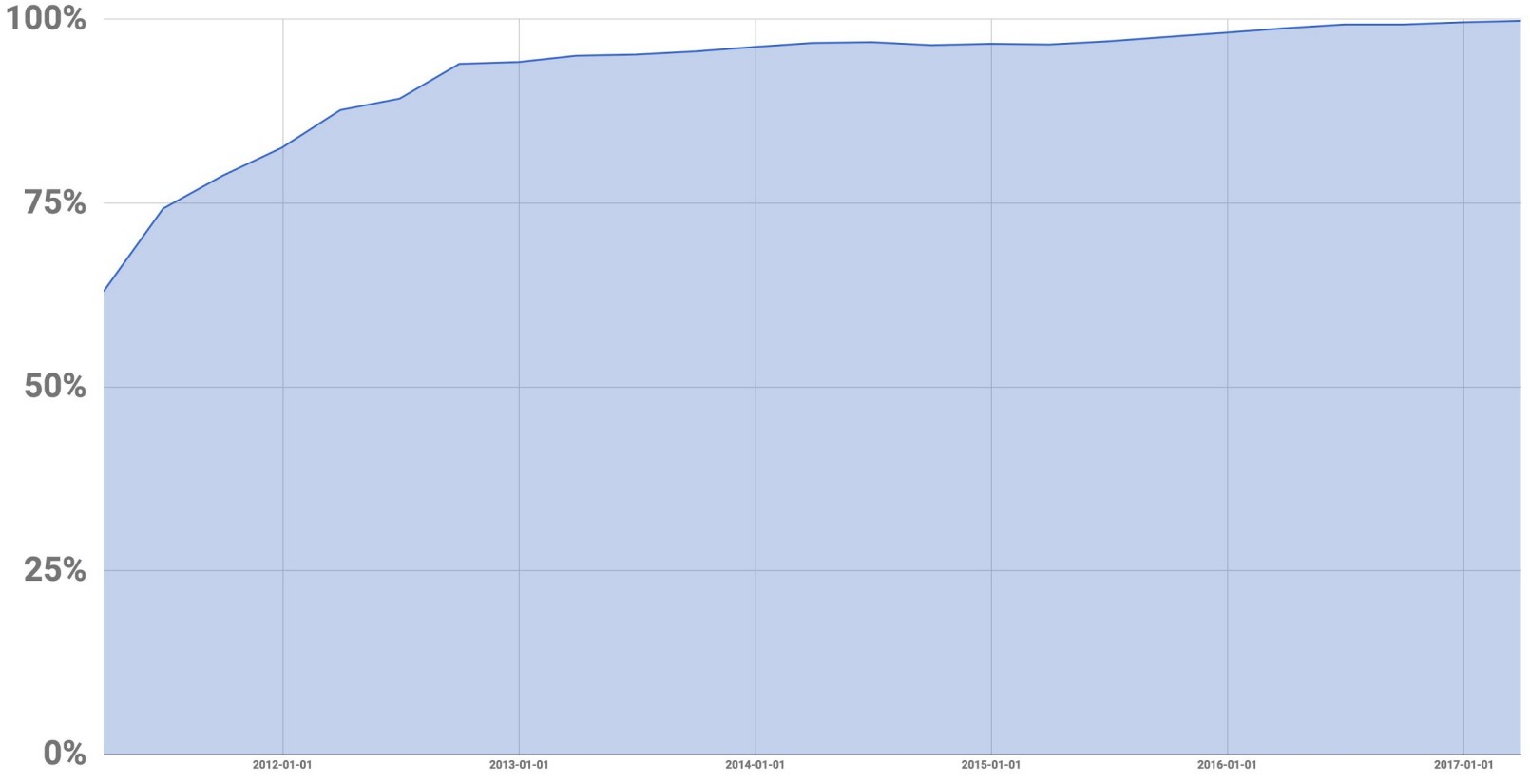
BTW, Unicode looking for contractors: FE & Performance



Core Code Libraries (ICU)

Unicode text handling	Formatting
Charset	Date & time
Conversions (200+)	Durations, intervals
Detection	Messages
Collation & Searching	Numbers, currencies
Resource Bundles	Measurement units
Calendar & Time Zones	Plurals
Unicode Regular Expressions	Transforms
Boundaries: word, line, …	Normalization
	Casing
	Transliterations
…	

ICU on Smartphones, Market Share

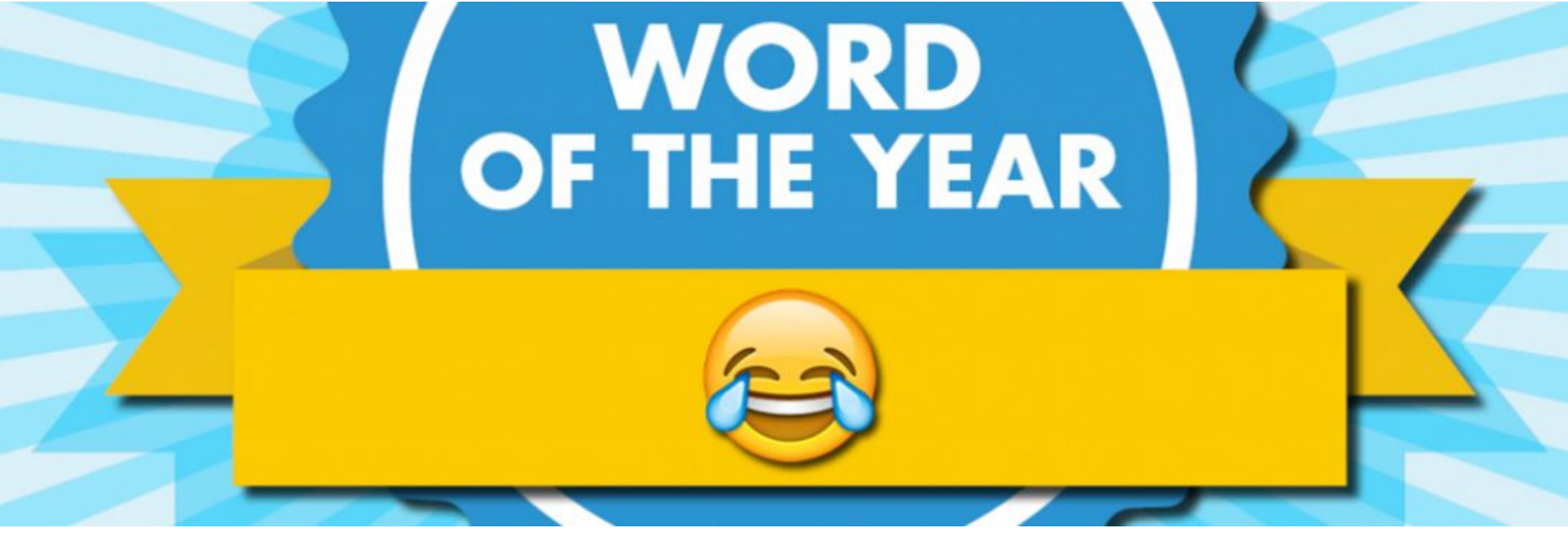


IDC: Worldwide smartphone shipments — 2011Q1..2017Q1

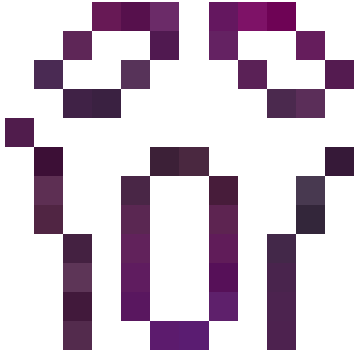
In a nutshell

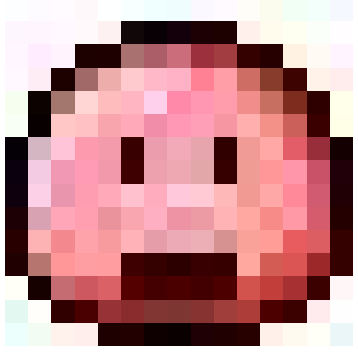
- Mature, widely used set of C/C++ and Java libraries
- Identical results on all platforms/programming languages
  - C/C++ (ICU4C): 30+ platforms/compilers
  - Java (ICU4J): Oracle, IBM JRE, Android
  - PyICU & other wrappers
- Customizable & Modular

Oxford Dictionaries



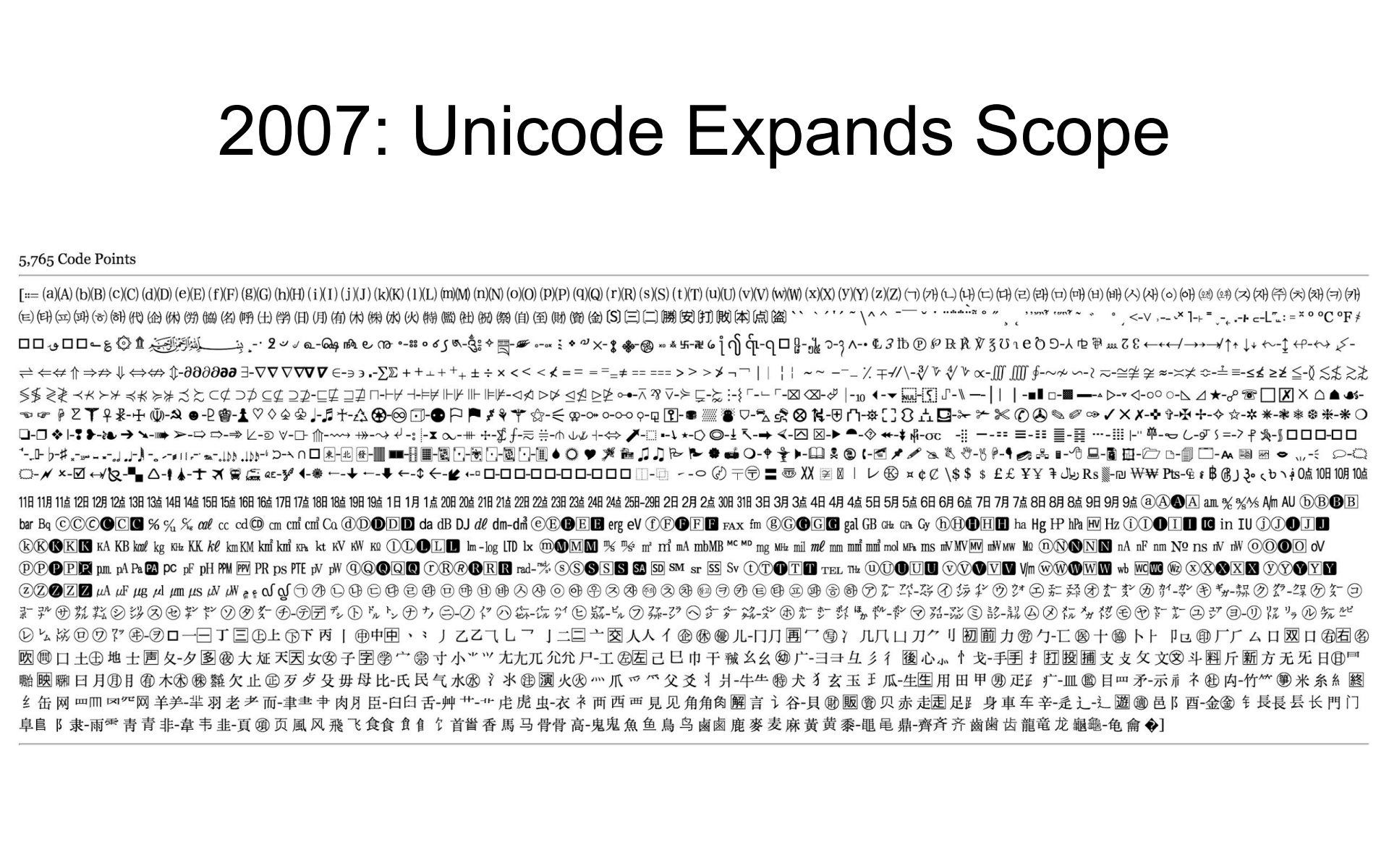
1999: Japanese Emoji



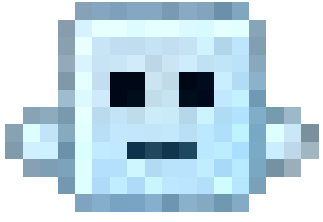


http://www.unicode.org/reports/tr51/#Introduction

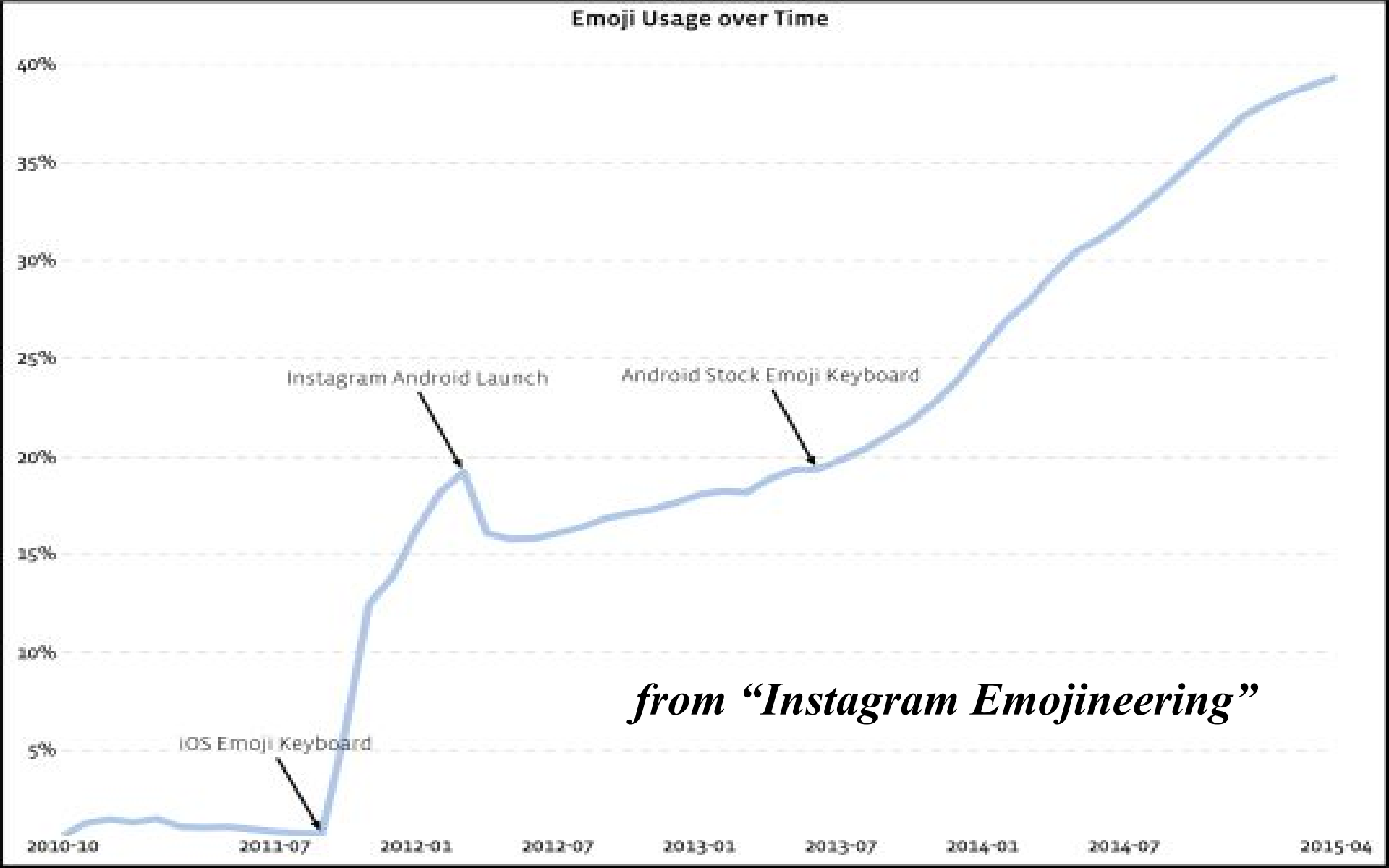
2007: Unicode Expands Scope



2008: Gmail / iPhone / …



2010: Unicode emoji


Emoji Usage over Time
40%
35%
30%
25%
20%
15%
10%
5%
Instagram Android Launch
Android Stock Emoji Keyboard
iOS Emoji Keyboard
2010-10
2011-07
2012-01
2012-07
2013-01
2013-07
2014-01
2014-07
2015-04
from "Instagram Emojineering"

2017

U+1F631

Details: unicode.org/emoji/charts

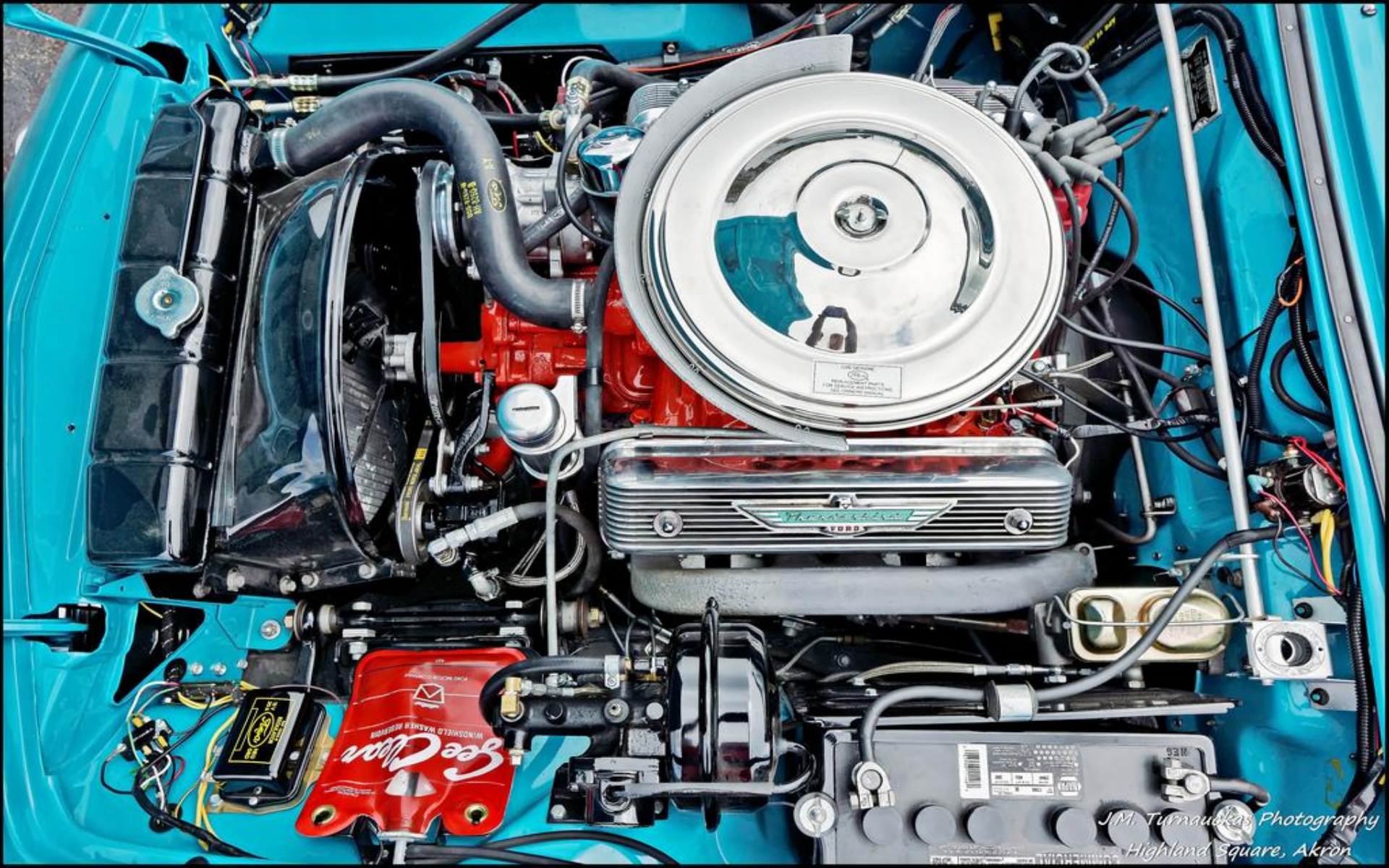
Emoji By-Product

MySQL 8.0 RC1 Now Available

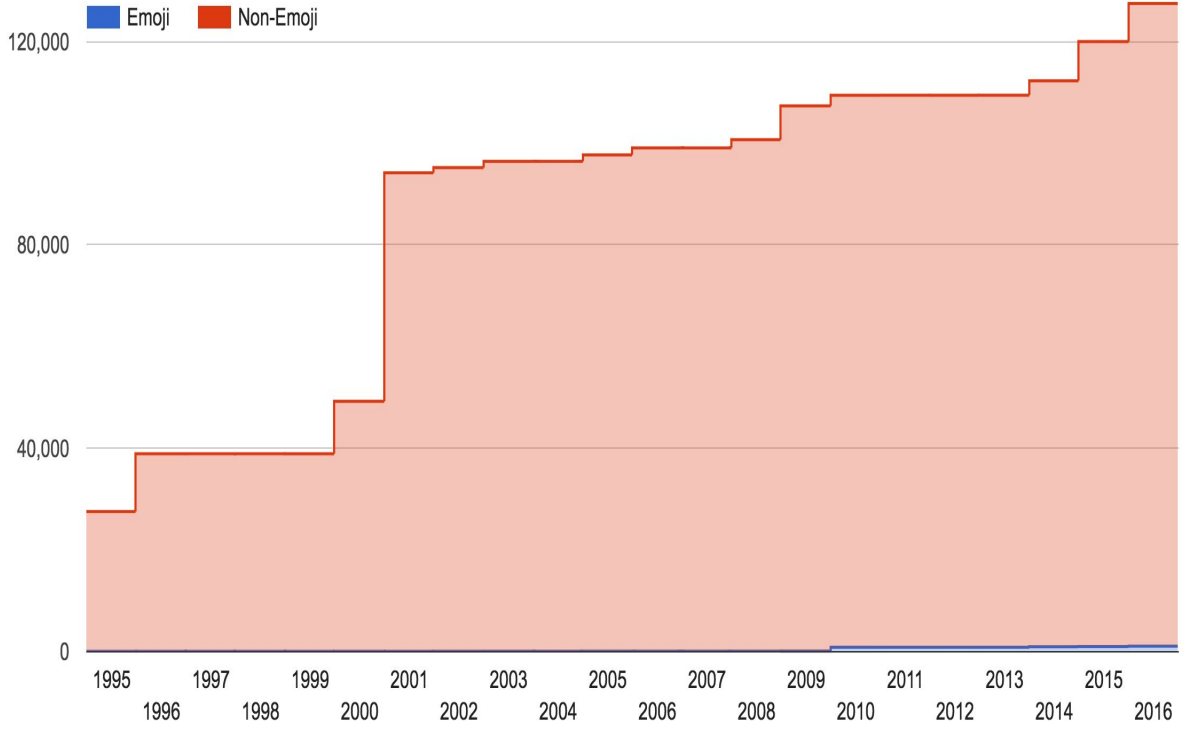
Sep 27, 2017

. . .

In addition, he adds, "Unicode (or more specifically UTF-8 encoding) has become universal even in English speaking markets. A key driver is mobile applications, where emojis are frequently used as character input. To support modern applications, it is important to have first-class support for UTF-8 out of the box."


FORD
See Clear
WINDSHIELD WASHER RESERVOIR
J.M. Turnaudkas Photography
Highland Square, Akron

Growth of Unicode Characters



How many? (Emoji 5.0)

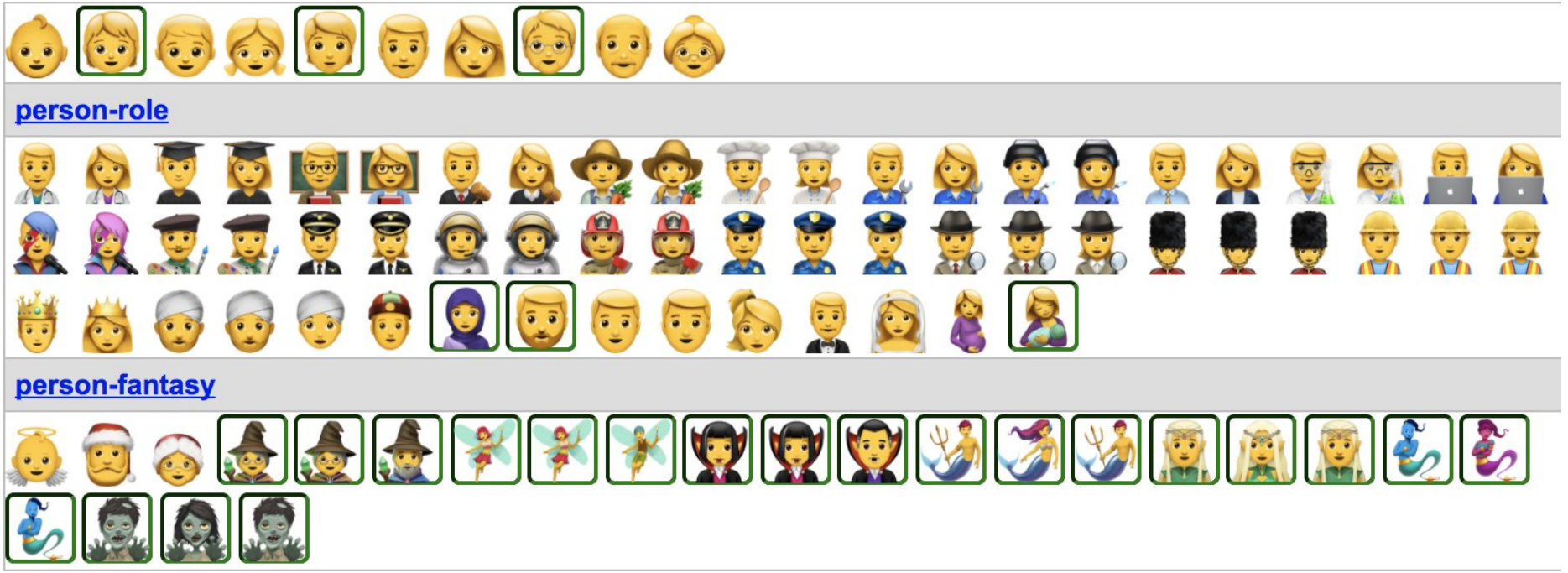
	Smileys & People	Animals & Nature	Food & Drink	Travel & Places	Activities	Objects	Symbols	Flags	Other	Total
character	297	113	102	207	60	162	193	5	0	**1,139**
keycap seq	0	0	0	0	0	0	12	0	0	**12**
flag seq	0	0	0	0	0	0	0	258	0	**258**
tag seq	0	0	0	0	0	0	0	3	0	**3**
mod seq	510	0	0	0	0	0	0	0	0	**510**
zwj seq + gender	43	0	0	0	0	0	0	0	0	**43**
zwj seq + modifier	160	0	0	0	0	0	0	0	0	**160**
zwj seq + gender, modifier	195	0	0	0	0	0	0	0	0	**195**
zwj seq other	61	0	0	0	0	0	0	1	0	**62**
Subtotal	**1,266**	**113**	**102**	**207**	**60**	**162**	**205**	**267**	**0**	**2,382**
component	5	0	0	0	0	0	0	0	38	**43**
typical dup	241	0	0	0	0	0	0	0	0	**241**
Total	**1,512**	**113**	**102**	**207**	**60**	**162**	**205**	**267**	**38**	**2,666**

Emoji Properties

Property	Description
Emoji	Emoji characters, also in sequences
Emoji_Presentation	Presented as emoji by default
Emoji_Modifier	Skin tones
Emoji_Modifier_Base	Bases for skin tones
Emoji_Component	For sequences, not typically on keyboard
Extended_Pictographic	CLDR: for proper (and future-proofed) segmentation.

Details: unicode.org/reports/tr51

Sort Order (CLDR)



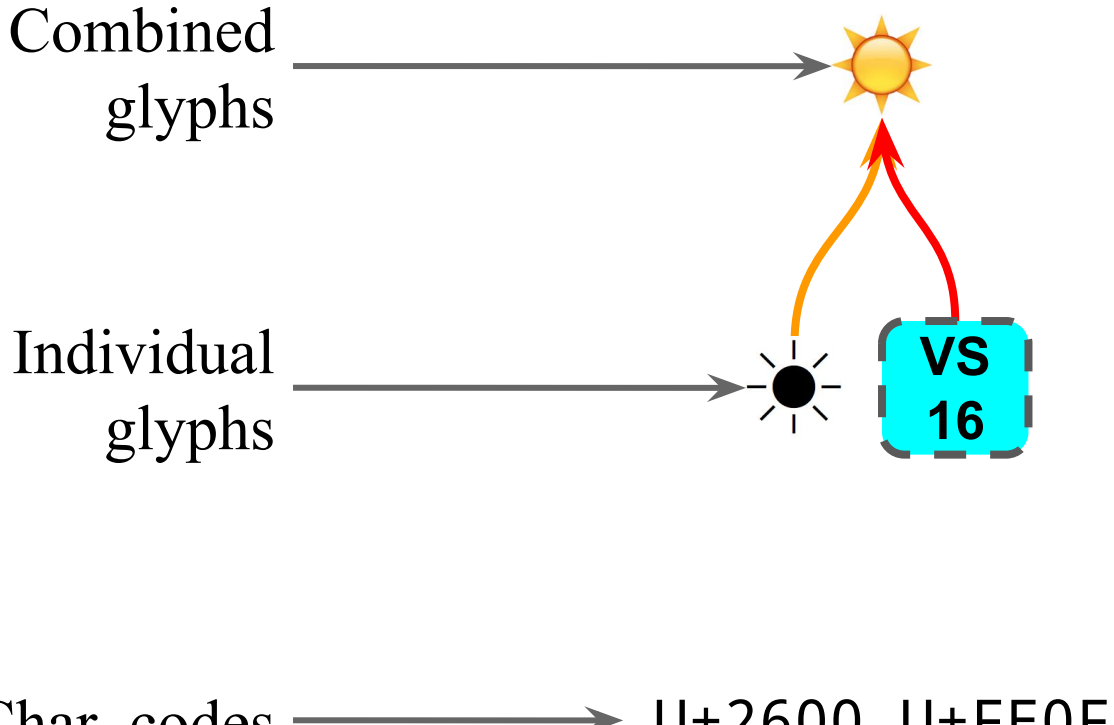
unicode.org/emoji/charts

Names, Keywords (CLDR)

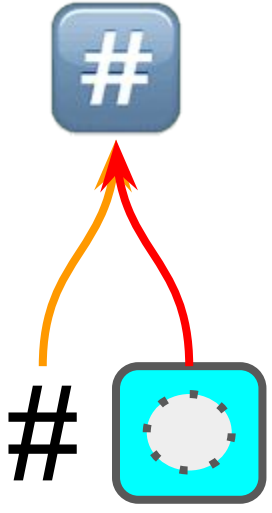
nerd face	nørdansigt	nerderig gezicht	Nerd-Smiley	nördaandlit
face | geek | nerd	ansigt | nørd	geek | gezicht | nerd	Gesicht | Nerd	andlit | lúði | nörd

unicode.org/cldr/charts/32/annotations

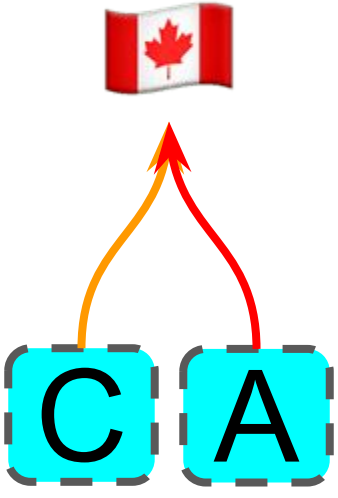
Variation


Combined glyphs
Individual glyphs
VS 16
Char. codes
U+2600 U+FE0F

Keycaps



Flags



Ill Formed

C

Invalid (but well-formed)

C B

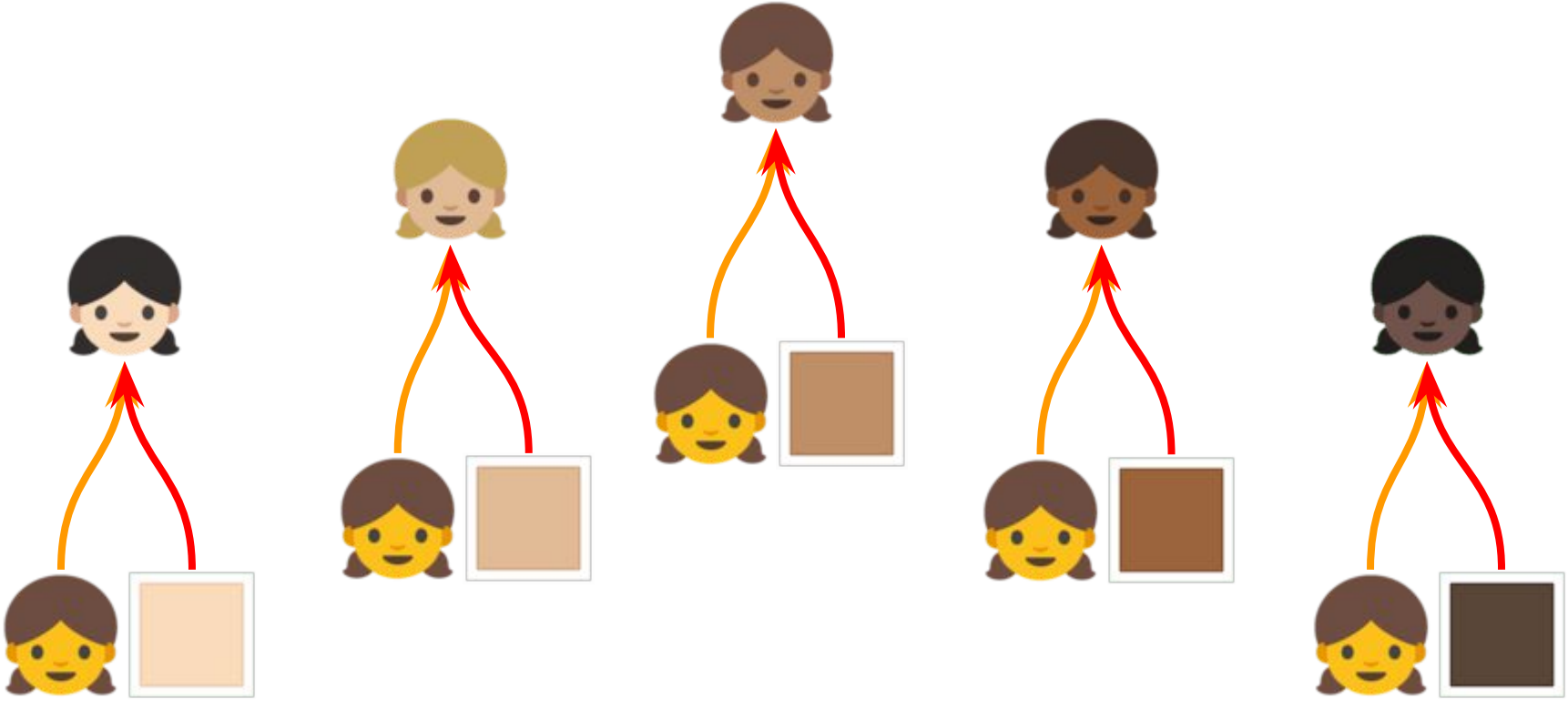
Valid (but not RGI)

C S

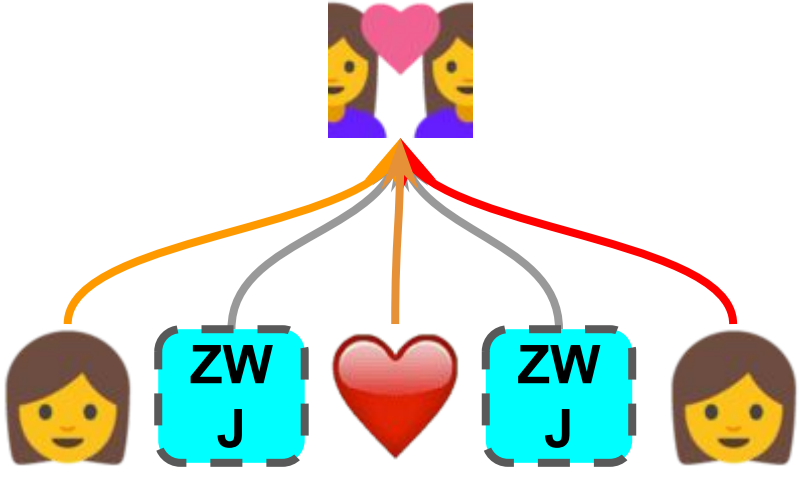
RGI

C A

Skin Tones

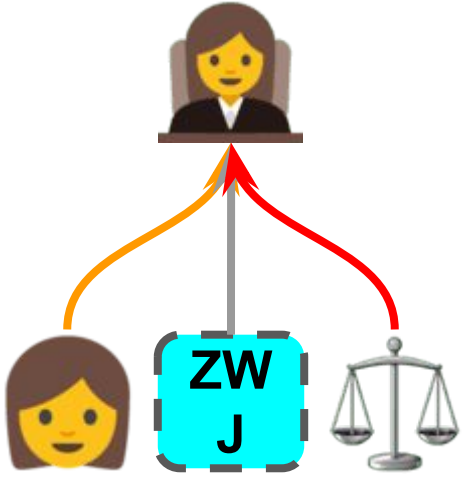


Emoji ZWJ Sequence

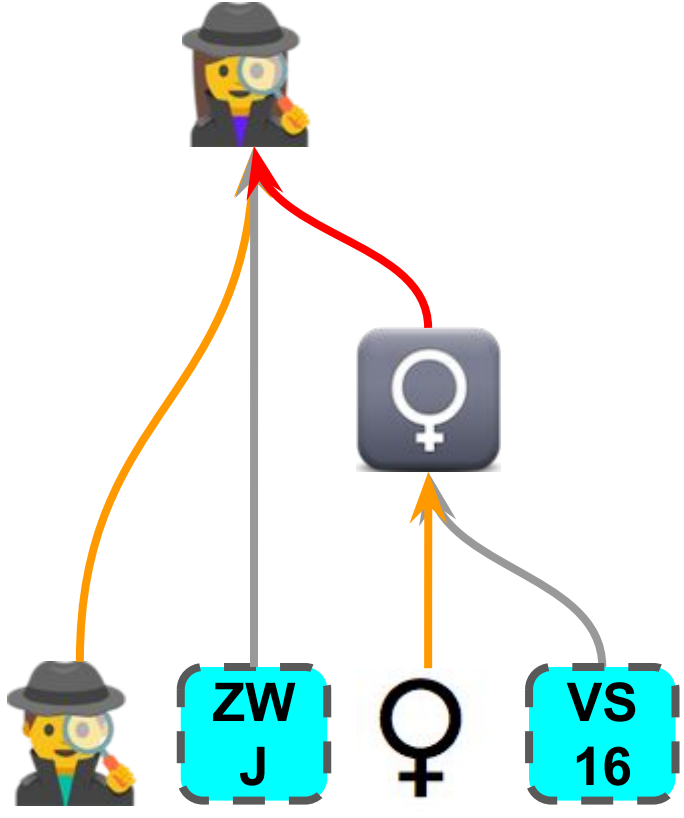



Stephen Colbert

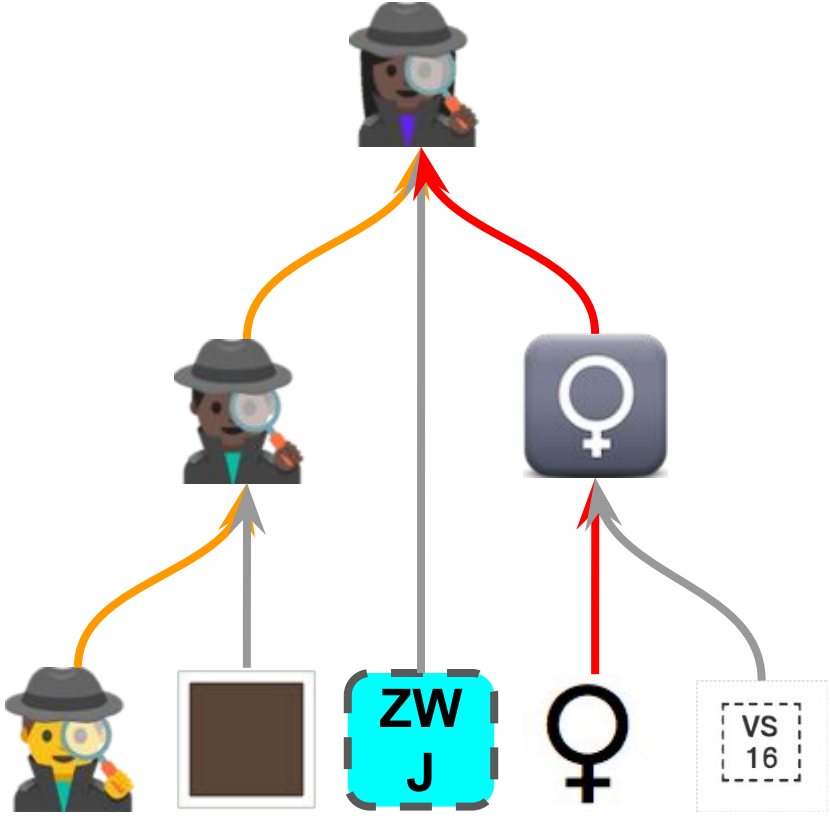
Gendered w/ Object



Gendered w/ Sign

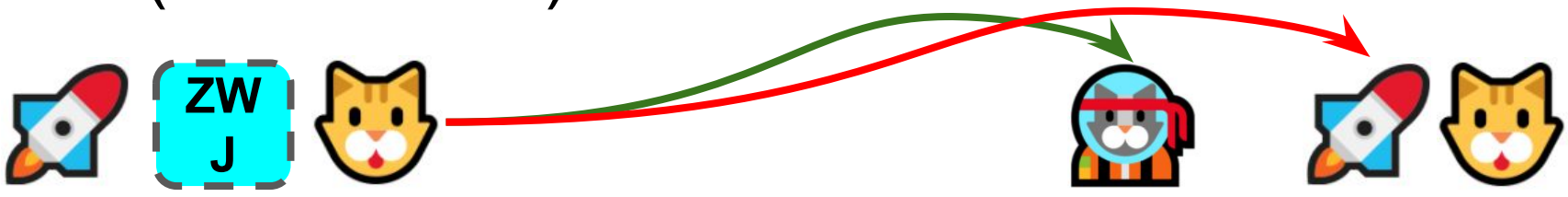


Gendered w/ Skintone & Sign



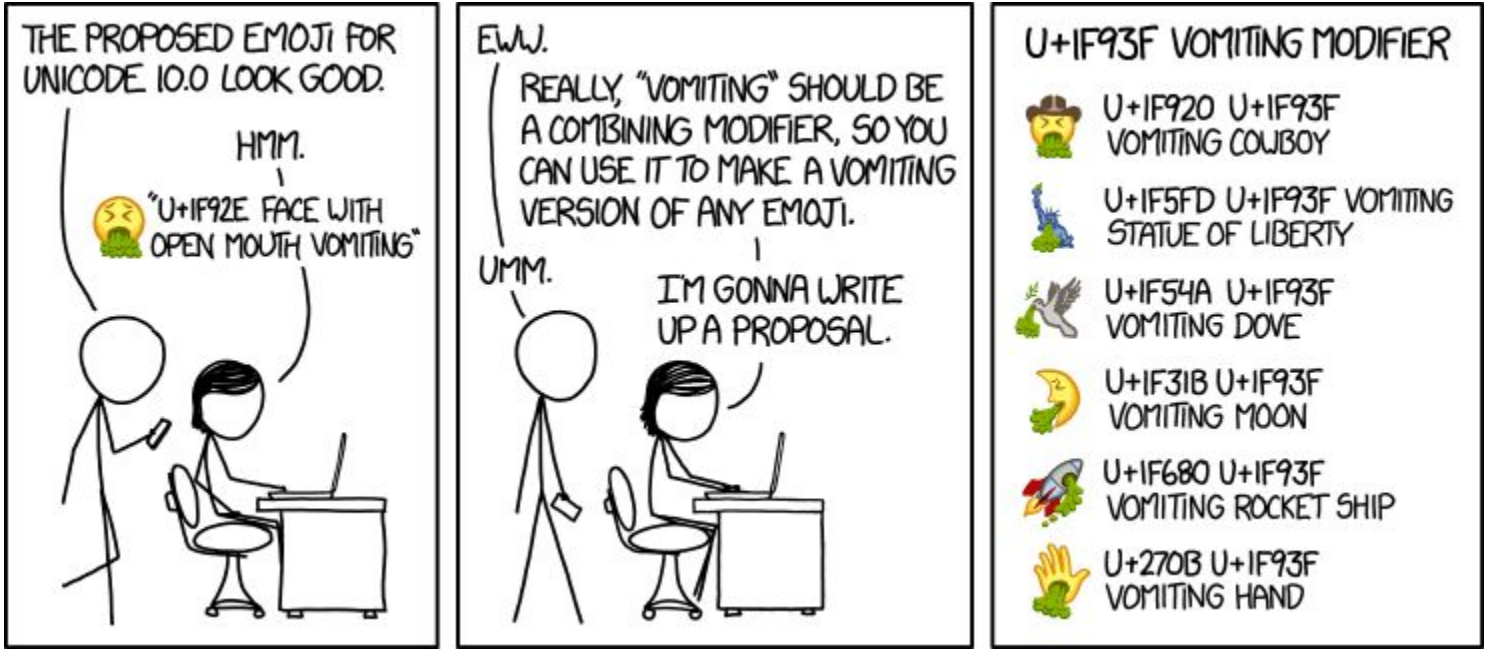
Ill Formed

Valid (but not RGI)



RGI

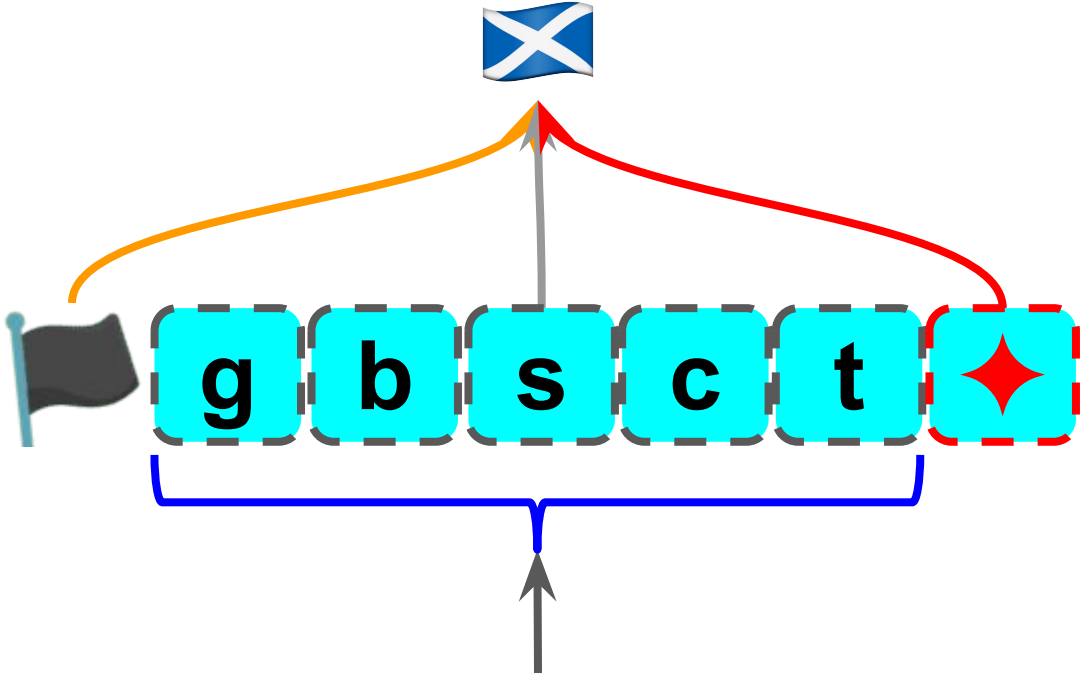
“Valid but not RGI?



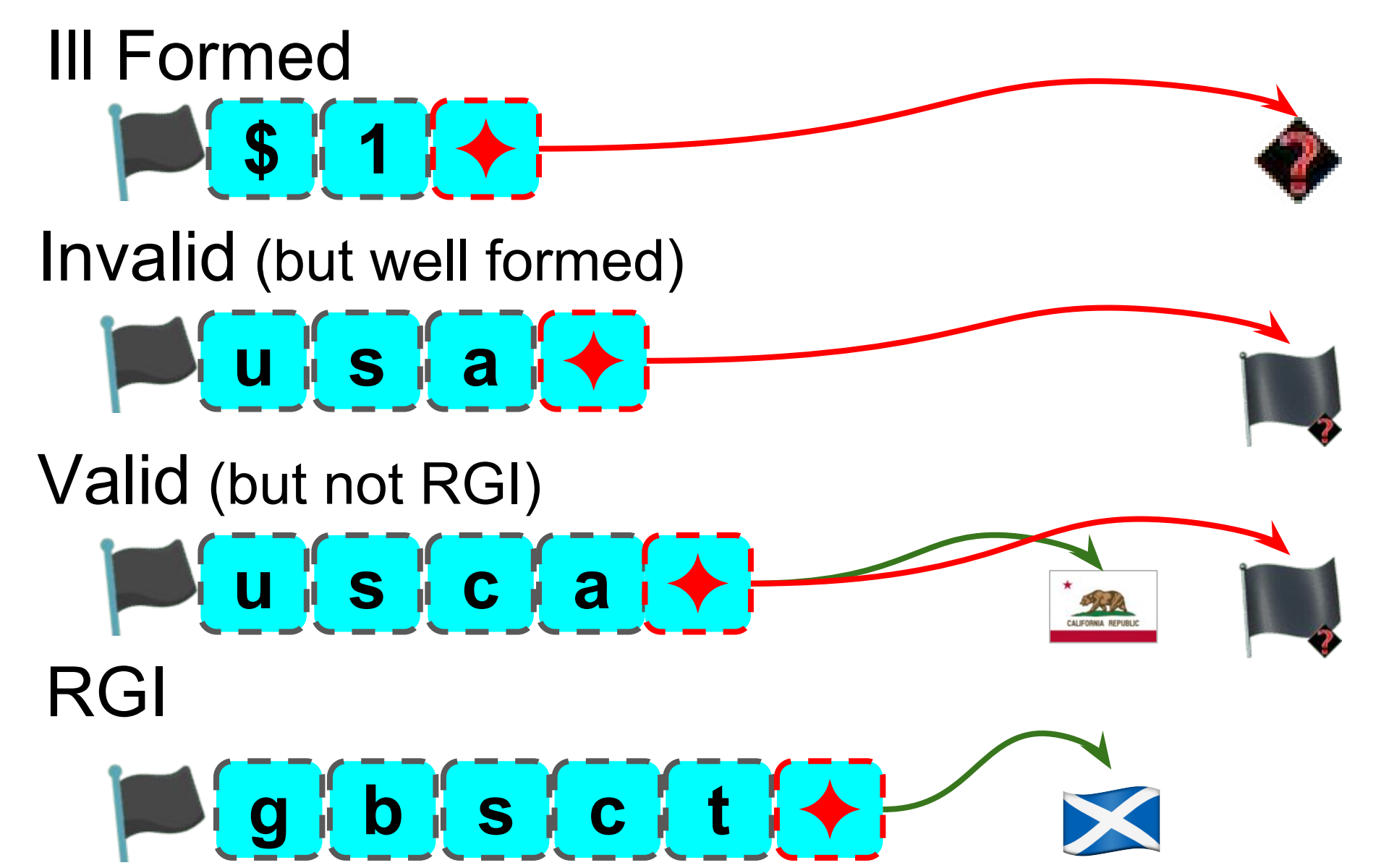
See https://xkcd.com/1813/

“Modifier” not needed: **Cowboy + ZWJ + Vomiting Face**

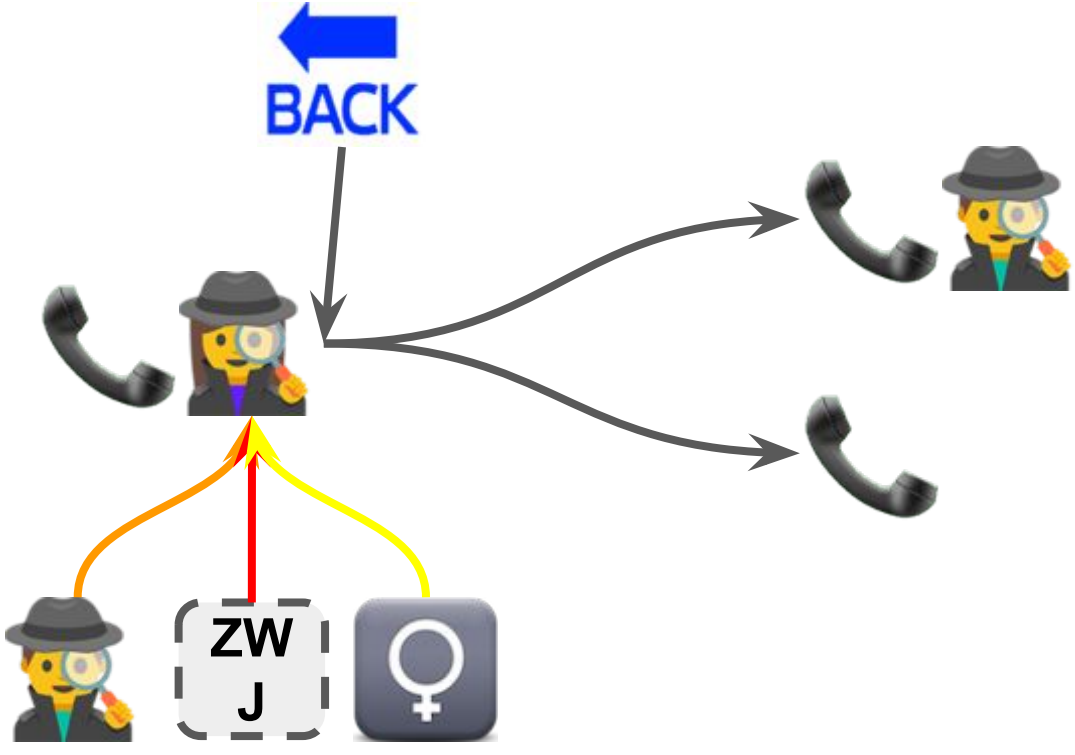
Tag Sequences



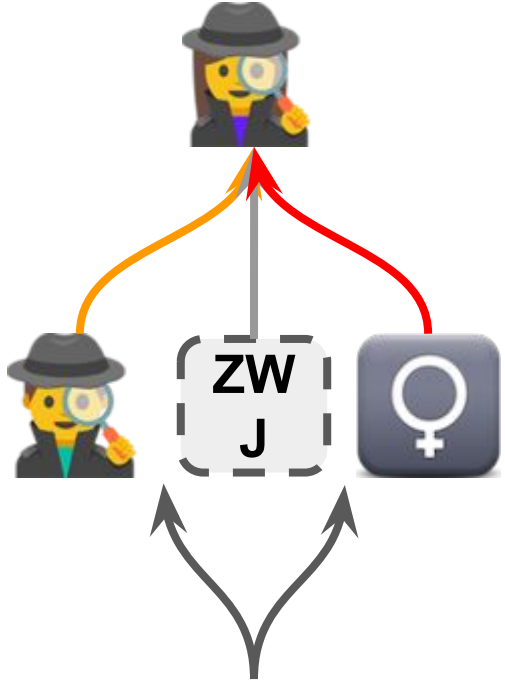
Subdivision = Province, State, Canton,...


Ill Formed
$ 1
Invalid (but well formed)
u s a
Valid (but not RGI)
u s c a
CALIFORNIA REPUBLIC
RGI
g b s c t

UI Actions: Backspace Example

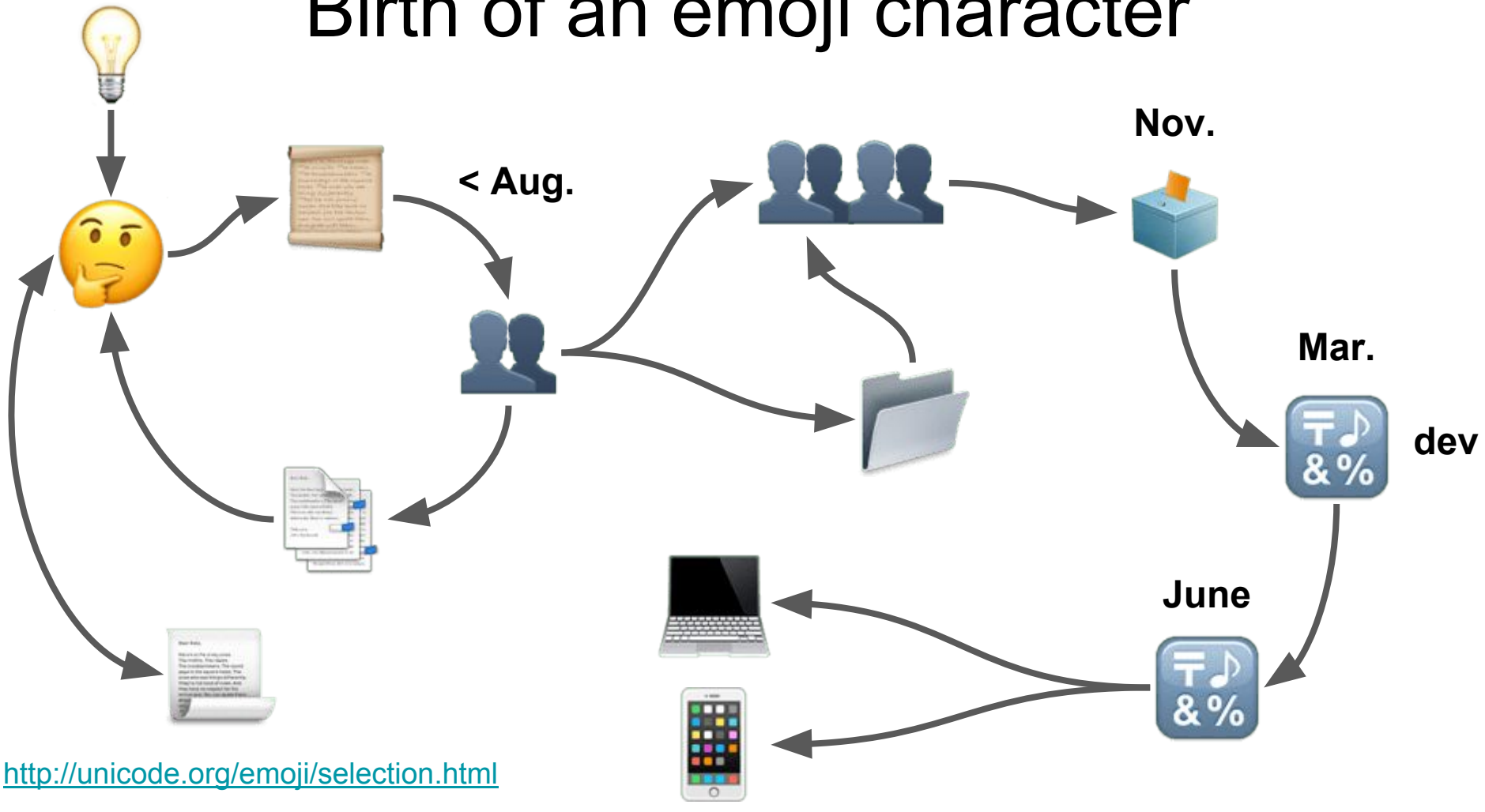


Segmentation



No linebreak here!

Birth of an emoji character



http://unicode.org/emoji/selection.html

Selection Factors

Factors for Inclusion +

A. Compatibility
B. Expected usage level
C. Image distinctiveness
D. Completeness
E. Frequently requested

Factors for Exclusion −

F. Overly specific
G. Open-ended
H. Already Representable
I. Logos, brands, … specific people, deities
J. Transient
K. Like compatibility emoji

http://unicode.org/emoji/selection.html

Proposal for a New Emoji: Cricket

Susan Burtner
<susan.burtner@gmail.com>

September 15, 2016

Abstract

The intention of this proposal is to advocate for the inclusion of emoji, the cricket, as a Unicode emoji character. The symbolic signi of a cricket is multicultural. To some it can be viewed as a sym illness, while to others it can be seen as a symbol of prosperity. In forms of American media, the cricket, particularly the *sound* of a chirping, implies an awkward silence. None of the current emo replace the varied symbolism implied by the cricket, or express th of "awkwardness" as clearly as the cricket does. The cricket emoj current design captures all of these meanings while still being ve enough to stand for a normal cricket.

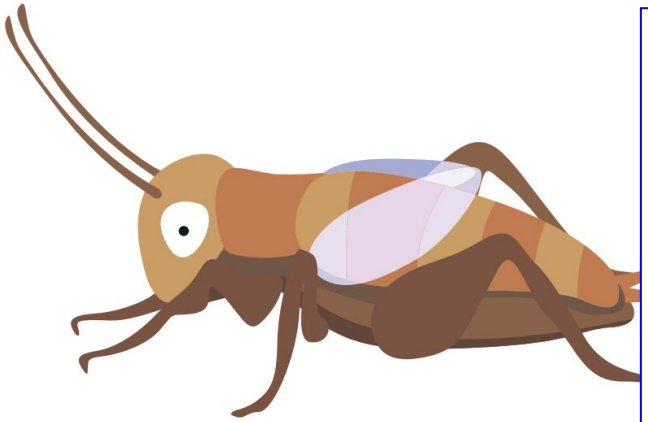


Figure 1: *A cricket created by Susan Burtner and Yoh Rosen, free for use and distribution by the Unicode Conso tium*

1 Introduction

In many cultures around the world, the sound and sight of a cricket a variety of good and bad prospects. In Brazil, a cricket can be se

2 Factors for Inclusion

2.1 Compatibility

Unfortunately, no emoji in the Unicode or any other system has created a cricket emoji like this. While this emoji cannot speak to the completeness and compatibility of other systems, it represents a unique opportunity to begin emoji representation for the ideas of "prosperity", "luck", and/or "awkwardness".

[1] "The Folklore and Mythology Surrounding Crickets": *http://www.crystalwi magical/legends-fables-and-lore/folklore/the-folklore-and-mythology-surroun*
[2] "Chinese Cricket Culture": *http://www.insects.org/ced3/chinese_crcul.html*
[3] "Native American Cricket Mythology": *http://www.universalsky.com/Art cricket_intuition_power_belief.htm*

2.2 Expected usage level

2.2.1 Frequency

It is the author's expectation that the cricket emoji will enjoy wide As mentioned above, the cricket can symbolize a variety of things, to prosperity as well as happiness from its singing presence. The long been a sign for things to come, and in conjunction with other e symbolize how one perceives a future outcome. In the same vein, a c could be used to wish a friend luck or good fortune for a future eve

2.2.2 Multiple usages

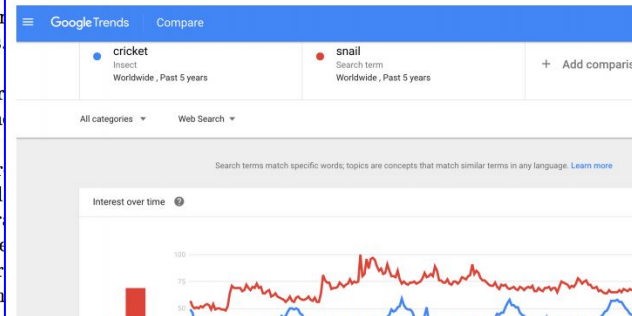
One of the greatest strengths of the cricket emoji is that it can stand for a creature with mythological significance, an awkward situation, and cricket that you might see in your home. A particularly popular cricket symbol in American media is its expression of awkwardness.

The cricket emoji could be used in situations in which no words or in combination or otherwise, suggest that very common and distin awkwardness. A search of "cricket" among the Unicode Code Chart one result: the cricket bat and ball. However, there are multiple r the popular association of crickets and awkwardness online. Several GIFs use the text "*crickets*" or "*cricket* *cricket*" to demonstr of dialogue in the image. A Google search for "crickets meme" yie results mostly pertaining to this idea of awkwardness. When sear ward silence sound effect" in YouTube, 13 of the 20 results on th

2.3 Image distinctiveness

People often confuse the cricket and the grasshopper. Some people one or the other makes noises, and they both do, in fact. The a chirping sound with its wings, and the grasshopper makes a sim sound with its legs.[4] Fortunately for this proposal, neither the gra the cricket exists as an emoji. The other emoji that come close, ant, honeybee, or lady beetle, could still not be used as suitable for the cricket. A comparable emoji is the snail, since it is also a a more distinct shape and bodily characteristics. When the cricket to a snail in a Google Search it can be seen that the two have dec popularity. A cricket search in relation to a snail search has a 3 worldwide over the past five years (about 56.52%) but the cricket is somewhat of an upward trend. When compared to a spider (ex within the same time frame this ratio is 2/40 or about 5%, and co lady beetle is 45/1 or about 4500%.

[4] "Grasshoppers and Crickets (Order: Orthoptera)": *https://www.amentsoc.o files/orders/orthoptera.html*

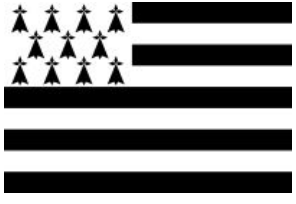


Emoji Candidates (click here)

Emoji 6.0 - Sample Requests

- Emojification

- RGI Tag Sequence



- RGI ZWJ Sequence

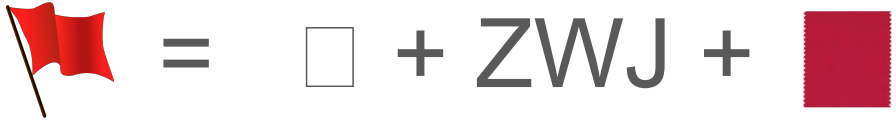
🏴‍☠️

1F3F4 + ZWJ + 2620

Issues

Gender Strategy — Complete Gender-Neutral forms

Color Swatches



Directions

Appears to User	Internal Representation	Fallback Appearance
🏃	🏃 ZWJ ⬅️	🏃 ⬅️
🏃	🏃 ZWJ ➡️	🏃 ➡️

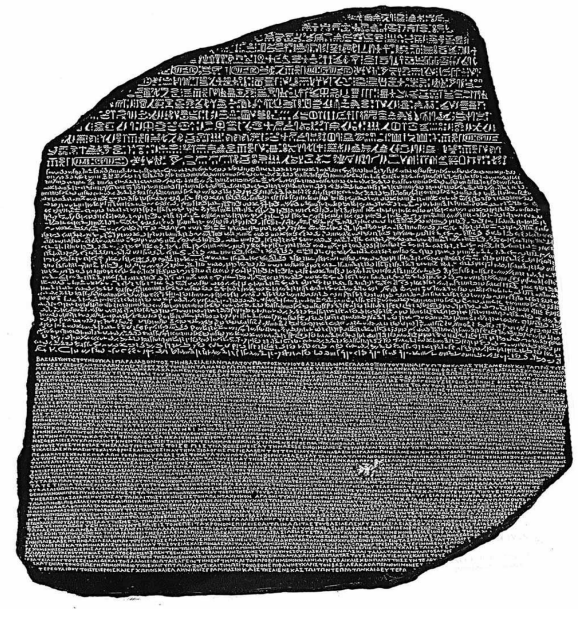
List of Proposals (click here)

A Greater Mission

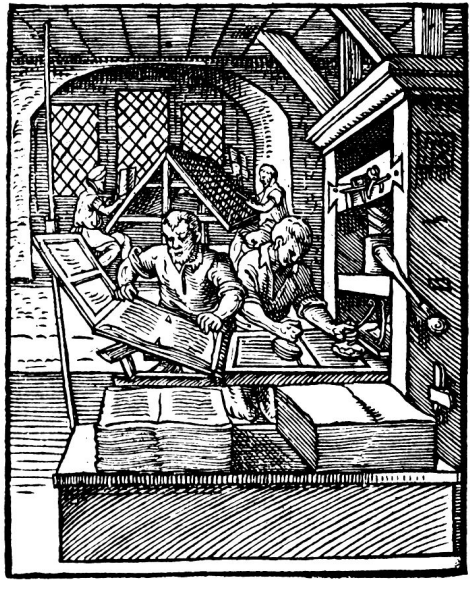
- Already vital to language support on computers:

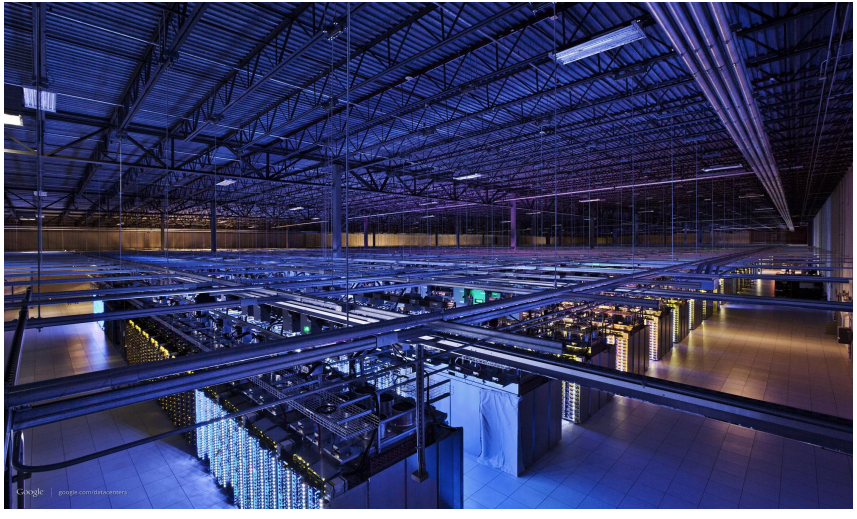
Every keystroke, every font, every language on every computer uses Unicode

- Vision: “No language left behind”
    - Existing membership focused on ‘short tail’ (more commercially relevant)
    - Funding for ‘long tail’ languages/characters (culturally significant)

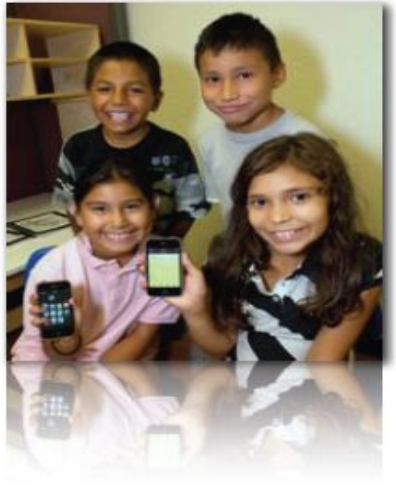


Preserving Cultural Heritage



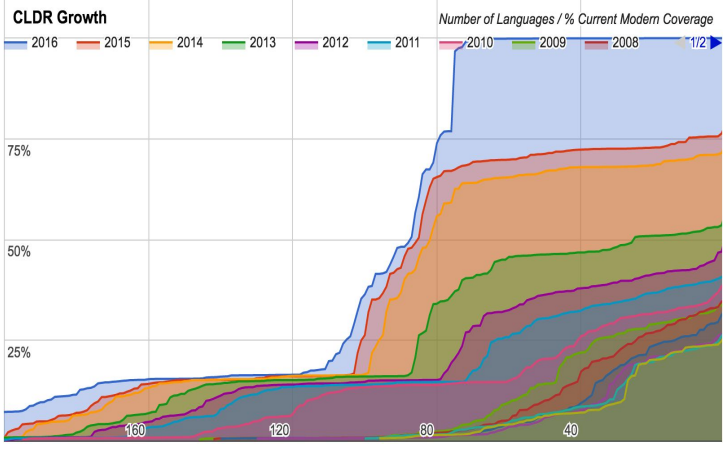


Reduce Digital Divide



Benefiting *Digitally-Disadvantaged Languages*

Emoji pushed products to improve their Unicode handling, but we have a long way to go in supporting more languages.



So we created a fundraising program for DDLs…

S&A Group

http://www.unicode.org/consortium/adopt-a-character.html

Questions?
(aka AMA)

Unicode Consortium